(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
9 August 2001 (09.08.2001)

PCT

(10) International Publication Number
WO 01/57273 A2

(51) International Patent Classification[7]: C12Q 1/68, G06F 19/00, C07K 14/47

(21) International Application Number: PCT/US01/00664

(22) International Filing Date: 30 January 2001 (30.01.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

| | | |
|---|---|---|
| 60/180,312 | 4 February 2000 (04.02.2000) | US |
| 60/207,456 | 26 May 2000 (26.05.2000) | US |
| 09/608,408 | 30 June 2000 (30.06.2000) | US |
| 09/632,366 | 3 August 2000 (03.08.2000) | US |
| 60/234,687 | 21 September 2000 (21.09.2000) | US |
| 60/236,359 | 27 September 2000 (27.09.2000) | US |
| 0024263.6 | 4 October 2000 (04.10.2000) | GB |

(71) Applicant *(for all designated States except US)*: **MOLECULAR DYNAMICS, INC.** [US/US]; 928 East Arques Avenue, Sunnyvale, CA 94086 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **PENN, Sharron, G.** [GB/US]; 617 South Delaware Street, San Mateo, CA 94402 (US). **HANZEL, David, K.** [US/US]; 988 Loma Verde Avenue, Palo Alto, CA 94303 (US). **CHEN, Wensheng** [CN/US]; 210 Easy Street #25, Mountain View, CA 94043 (US). **RANK, David, R.** [US/US]; 117 El Dorado Commons, Fremont, CA 94539 (US).

(74) Agent: **RONNING, Royal, N., Jr.**; Amersham Pharmacia Biotech, Inc., 800 Centennial Avenue, Piscataway, NJ 08855 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

**Published:**

— *without international search report and to be republished upon receipt of that report*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

WO 01/57273 A2

(54) Title: HUMAN GENOME-DERIVED SINGLE EXON NUCLEIC ACID PROBES USEFUL FOR ANALYSIS OF GENE EXPRESSION IN HUMAN ADULT LIVER

(57) Abstract: A single exon nucleic acid microarray comprising a plurality of single exon nucleic acid probes for measuring gene expression in a sample derived from human adult liver is described. Also described are single exon nucleic acid probes expressed in the adult liver and their use in methods for detecting gene expression.

# HUMAN GENOME-DERIVED SINGLE EXON NUCLEIC ACID PROBES USEFUL FOR ANALYSIS OF GENE EXPRESSION IN HUMAN ADULT LIVER

## CROSS REFERENCE TO RELATED APPLICATIONS

The present application is a continuation-in-part of U.S. patent application serial nos. 09/632,366, filed August 3, 2000 and 09/608,408, filed June 30, 2000; claims the benefit under 35 U.S.C. s 119(e) of U.S.provisional patent application serial nos. 60/236,359, filed September 27, 2000, 60/234,687, filed September 21, 2000, 60/207,456, filed May 26, 2000, and 60/180,312, filed February 4, 2000; and further claims the benefit under 35 U.S.C. s 119(a) of UK patent application no. 0024263.6, filed October 4, 2000, the disclosures of which are incorporated herein by reference in their entireties.

## REFERENCE TO SEQUENCE LISTING AND INCORPORATION BY REFERENCE THEREOF

The present application includes a Sequence Listing in electronic format, filed pursuant to PCT Administrative Instructions 801 - 806 on a single CD-R disc, in triplicate, containing a file named pto_ADULT_LIVER.txt, created 24 January 2001, having 26,335,065 bytes. The Sequence Listing contained in said file on said disc is incorporated herein by reference in its entirety.

## Field of the Invention

The present invention relates to genome-derived single exon microarrays useful for verifying the expression of regions of genomic DNA predicted to encode protein. In particular, the present invention relates to unique genome-derived single exon nucleic acid probes expressed in human

adult liver and single exon nucleic acid microarrays that include such probes.

Background of the Invention

For almost two decades following the invention of general techniques for nucleic acid sequencing, Sanger *et al.*, *Proc. Natl. Acad. Sci. USA* 70(4):1209-13 (1973); Gilbert *et al.*, *Proc. Natl. Acad. Sci. USA* 70(12):3581-4 (1973), these techniques were used principally as tools to further the understanding of proteins – known or suspected – about which a basic foundation of biological knowledge had already been built. In many cases, the cloning effort that preceded sequence identification had been both informed and directed by that antecedent biological understanding.

For example, the cloning of the T cell receptor for antigen was predicated upon its known or suspected cell type-specific expression, by its suspected membrane association, and by the predicted assembly of its gene via T cell-specific somatic recombination. Subsequent sequencing efforts at once confirmed and extended understanding of this family of proteins. Hedrick *et al.*, *Nature* 308(5955):153-8 (1984).

More recently, however, the development of high throughput sequencing methods and devices, in concert with large public and private undertakings to sequence the human and other genomes, has altered this investigational paradigm: today, sequence information often precedes understanding of the basic biology of the encoded protein product.

One of the approaches to large-scale sequencing is predicated upon the proposition that expressed sequences – that is, those accessible through isolation of mRNA – are of greatest initial interest. This "expressed sequence tag" ("EST") approach has already yielded vast

amounts of sequence data (see for example Adams *et al.*, *Science* 252:1651 (1991); Williamson, *Drug Discov. Today* 4:115 (1999)). For nucleic acids sequenced by this approach, often the only biological information that is known *a priori* with any certainty is the likelihood of biologic expression itself. By virtue of the species and tissue from which the mRNA had originally been obtained, most such sequences are also annotated with the identity of the species and at least one tissue in which expression appears likely.

More recently, the pace of genomic sequencing has accelerated dramatically. When genomic DNA serves as the initial substrate for sequencing efforts, expression cannot be presumed; often the only *a priori* biological information about the sequence includes the species and chromosome (and perhaps chromosomal map location) of origin.

With the ever-accelerating pace of sequence accumulation by directed, EST, and genomic sequencing approaches – and in particular, with the accumulation of sequence information from multiple genera, from multiple species within genera, and from multiple individuals within a species – there is an increasing need for methods that rapidly and effectively permit the functions of nucleic sequences to be elucidated. And as such functional information accumulates, there is a further need for methods of storing such functional information in meaningful and useful relationship to the sequence itself; that is, there is an increasing need for means and apparatus for annotating raw sequence data with known or predicted functional information.

Although the increase in the pace of genomic sequencing is due in large part to technological changes in sequencing strategies and instrumentation, Service, *Science* 280:995 (1998); Pennisi, *Science* 283: 1822-1823 (1999), there is an important functional motivation as well.

While it was understood that the EST approach would rarely be able to yield sequence information about the noncoding portions of the genome, it now also appears the EST approach is capable of capturing only a fraction of a genome's actual expression complexity.

For example, when the *C. elegans* genome was fully sequenced, gene prediction algorithms identified over 19,000 potential genes, of which only 7,000 had been found by EST sequencing. *C. elegans* Sequencing Consortium, *Science* 282:2012 (1998). Analogously, the recently completed sequence of chromosome 2 of *Arabidopsis* predicts over 4000 genes, Lin *et al.*, *Nature*, 402:761 (1999), of which only about 6% had previously been identified via EST sequencing efforts. Although the human genome has the greatest depth of EST coverage, it is still woefully short of surrendering all of its genes. One recent estimate suggests that the human genome contains more than 146,000 genes, which would at this point leave greater than half of the genes undiscovered. It is now predicted that many genes, perhaps 20 to 50%, will only be found by genomic sequencing.

There is, therefore, a need for methods that permit the functional regions of genomic sequence — and most importantly, but not exclusively, regions that function to encode genes — to be identified.

Much of the coding sequence of the human genome is not homologous to known genes, making detection of open reading frames ("ORFs") and predictions of gene function difficult. Computational methods exist for predicting coding regions in eukaryotic genomes. Gene prediction programs such as GRAIL and GRAIL II, Uberbacher *et al.*, *Proc. Natl. Acad. Sci. USA* 88(24):11261-5 (1991); Xu *et al.*, *Genet. Eng.* 16:241-53 (1994); Uberbacher *et al.*, *Methods Enzymol.* 266:259-81 (1996); GENEFINDER, Solovyev *et al.*, *Nucl. Acids. Res.* 22:5156-63 (1994); Solovyev *et al.*,

*Ismb* 5:294-302 (1997); and GENESCAN, Burge *et al.*, *J. Mol. Biol.* 268:78-94 (1997), predict many putative genes without known homology or function. Such programs are known, however, to give high false positive rates. Burset *et al.*, *Genomics* 34:353-367 (1996). Using a consensus obtained by a plurality of such programs is known to increase the reliability of calling exons from genomic sequence. Ansari-Lari *et al.*, *Genome Res.* 8(1):29-40 (1998)

Identification of functional genes from genomic data remains, however, an imperfect art. For example, in reporting the full sequence of human chromosome 21, the Chromosome 21 Mapping and Sequencing Consortium reports that prior bioinformatic estimates of human gene number may need to be revised substantially downwards. *Nature* 405:311-199 (2000); Reeves, *Nature* 405:283-284 (2000).

Thus, there is a need for methods and apparatus that permit the functions of the regions identified bioinformatically — and specifically, that permit the expression of regions predicted to encode protein — readily to be confirmed experimentally.

Recently, the development of nucleic acid microarrays has made possible the automated and highly parallel measurement of gene expression. *Reviewed in* Schena (ed.), DNA Microarrays : A Practical Approach (Practical Approach Series), Oxford University Press (1999) (ISBN: 0199637768); *Nature Genet.* 21(1)(suppl):1 - 60 (1999); Schena (ed.), Microarray Biochip: Tools and Technology, Eaton Publishing Company/BioTechniques Books Division (2000) (ISBN: 1881299376).

It is common for microarrays to be derived from cDNA/EST libraries, either from those previously described in the literature, such as those from the I.M.A.G.E. consortium, Lennon *et al.*, *Genomics* 33(1):151-2 (1996), or from the construction of "problem specific" libraries targeted at a particular biological question, R.S. Thomas

*et al.*, *Cancer Res.* (in press). Such microarrays by definition can measure expression only of those genes found in EST libraries, and thus have not been useful as probes for genes discovered solely by genomic sequencing.

The utility of using whole genome nucleic acid microarrays to answer certain biological questions has been demonstrated for the yeast *Saccharomyces cerevisiae*. De Risi *et al.*, *Science* 278:680 (1997). The vast majority of yeast nuclear genes, approximately 95% however, are single exon genes, *i.e.*, lack introns, Lopez *et al.*, *RNA* 5:1135-1137 (1999); Goffeau *et al.*, *Science* 274:563-67 (1996), permitting coding regions more readily to be identified. Whole genome nucleic acid microarrays have not generally been used to probe gene expression from more complex eukaryotic genomes, and in particular from those averaging more than one intron per gene.

Diseases of the liver are a significant cause of human morbidity and mortality. Increasingly, genetic factors are being found that contribute to predisposition, onset, and/or aggressiveness of most, if not all, of these diseases; although causative mutations in single genes have been identified for some, these disorders are believed for the most part to have polygenic etiologies. There is a need for methods and apparatus that permit prediction, diagnosis and prognosis of diseases of the liver particularly those diseases with polygenic etiologies.

## Summary of the Invention

The present invention solves these and other problems in the art by providing methods and apparatus for predicting, confirming, and displaying functional information derived from genomic sequence. The present invention also provides apparatus for verifying the expression of putative genes identified within genomic

sequence.

In particular, the invention provides novel genome-derived single exon nucleic acid microarrays useful for verifying the expression of putative genes identified within genomic sequence.

The present invention also provides compositions and kits for the ready production of nucleic acids identical in sequence to, or substantially identical in sequence to, probes on the genome-derived single exon microarrays of the present invention.

Accordingly, in a first aspect of the invention, there is provided a spatially-addressable set of single exon nucleic acid probes for measuring gene expression in a sample derived from human adult liver, comprising a plurality of single exon nucleic acid probes according to any one of the nucleotide sequences set out in SEQ ID NOs: 1 - 13,109or a complementary sequence, or a portion of such a sequence.

By plurality is meant at least two, suitably at least 20, most suitably at least 100, preferably at least 1000 and, most preferably, upto 5000.

In one embodiment of the first aspect, each of said plurality of probes is separately and addressably amplifiable.

In an alternative embodiment, each of said plurality of probes is separately and addressably isolatable from said plurality.

In a preferred embodiment, each of said plurality of probes is amplifiable using at least one common primer. Preferably, each of said plurality of probes is amplifiable using a first and a second common primer.

In yet another embodiment, said set of single exon nucleic acid probes comprises between 50 - 20,000 probes, for example, 50 - 5000.

Suitably, said set of single exon nucleic acid

probes comprises at least 50 - 1000 discrete single exon nucleic acid probes having a sequence as set out in any of SEQ ID NOS.: 1 - 25,995or a complimentary sequence, or a portion of such a sequence.

Preferably, the average length of the single exon nucleic acid probes is between 200 and 500 bp. It is preferred that the average length should be at least 200bp, suitably at least 250bp, most suitably at least 300bp, preferably at least 400bp and, most preferably, 500 bp.

In another embodiment, the single exon nucleic acid probes lack prokaryotic and bacteriophage vector sequence. It is preferred that at least 50%, suitably at least 60%, most suitably at least 70%, preferably at least 75%, more preferably at least 80, 85, 90, 95 or 99% of said single exon nucleic acid probes lack prokaryotic and bacteriophage vector sequence.

In another preferred embodiment, said single exon nucleic acid lack homopolymeric stretches of A or T. It is preferred that at least 50%, suitably at least 60%, most suitably at least 70%, preferably at least 75%, more preferably at least 80, 85, 90, 95 or 99% of said single exon nucleic acid probes lack homopolymeric stretches of A or T.

Preferably, a spatially-addressable set of single exon nucleic acid probes in accordance with the first aspect of the invention is is addressably disposed upon a substrate.

Suitable substrates include a filter membrane which may, preferably, be nitrocellulose or nylon. The nylon may preferably, be positively-charged. Other suitable substrates include glass, amorphous silicon, crystalline silicon, and plastic. Further suitable materials include polymethylacrylic, polyethylene, polypropylene, polyacrylate, polymethylmethacrylate, polyvinylchloride, polytetrafluoroethylene, polystyrene, polycarbonate,

polyacetal, polysulfone, celluloseacetate, cellulosenitrate, nitrocellulose, and mixtures thereof.

In a second aspect of the invention, there is provided a microarray comprising a spatially addressable set of single exon nucleic acid probes in accordance with the first aspect of the invention.

In one embodiment, a genome-derived single-exon microarray is packaged together with such an ordered set of amplifiable probes corresponding to the probes, or one or more subsets of probes, thereon. In alternative embodiments, the ordered set of amplifiable probes is packaged separately from the genome-derived single exon microarray.

In another aspect, the invention provides genome-derived single exon nucleic acid probes useful for gene expression analysis, and particularly for gene expression analysis by microarray. In particular embodiments of this aspect, the present invention provides human single-exon probes that include specifically-hybridizable fragments of SEQ ID Nos. 13,110 - 25,995, wherein the fragment hybridizes at high stringency to an expressed human gene. In particular embodiments, the invention provides single exon probes comprising SEQ ID Nos. 1 - 13,109.

Accordingly, in a third aspect of the invention, there is provided a single exon nucleic acid probe for measuring human gene expression in a sample derived from human adult liver which is a nucleic acid molecule comprising a nucleotide sequence as set out in any of SEQ ID NOs.: 1 - 13,109or a complementary sequence or a fragment thereof wherein said probe hybridizes at high stringency to a nucleic acid expressed in the human adult liver.

In one embodiment, a single exon nucleic acid probe in accordance with the third aspect comprises a nucleotide sequence as set out in any of SEQ ID NOs.:

13,110 - 25,995 or a complementary sequence or a fragment thereof.

In a fourth aspect of the invention, there is provided a single exon nucleic acid probe for measuring human gene expression in a sample derived from human adult liver which is a nucleic acid molecule having a sequence encoding a peptide comprising a peptide sequence as set out in any of SEQ ID NOs.: 25,996 - 38,578 or a complementary sequence or a fragment thereof wherein said probe hybridizes at high stringency to a nucleic acid expressed in the human adult liver.

Preferably, a single exon nucleic acid probe in accordance with the third or fourth aspects of the invention comprises between at least 15 and 50 contiguous nucleotides of said SEQ ID NO:. It is preferred that the single exon nucleic acid probe comprises at least 15, suitably at least 20, more suitably at least 25 or preferably at least 50 contiguous nucleotides of said SEQ ID NO:.

In another preferred embodiment, a single exon nucleic acid probe in accordance with the third or fourth aspects of the invention is between 3kb and 25kb in length. It is preferred that said probe is no more than 3kb, suitably no more than 5kb, more suitably no more than 10kb, preferably 15kb, more preferably 20kb or, most preferably, no more than 20kb in length.

Preferably, a single exon nucleic acid probe in accordance with either the fifth or sixth aspect of the invention is DNA, preferably single-stranded DNA, RNA or PNA.

In another embodiment of either the third or fourth aspect of the invention, a single exon nucleic acid probe is detectably labeled. Suitable detectable labels include a radionuclide, a fluorescent label or a first member of a specific binding pair. Suitable fluorescent

labels include dyes such as cyanine dyes, preferably Cy3 and Cy5 although other suitable dyes will be known to those skilled in the art.

In a particularly preferred embodiment, a single exon nucleic acid probe in accordance with either the third or fourth aspect of the invention lacks prokaryotic and bacteriophage vector sequence. In yet another embodiment, a single exon nucleic acid probe in accordance with either the third or fourth aspect of the invention lacks homopolymeric stretches of A or T.

In a fifth aspect of the invention, there is provided an amplifiable nucleic acid composition, comprising:

the single exon nucleic acid probe in accordance with either of the third or fourth aspects of the invention; and at least one nucleic acid primer;

wherein said at least one primer is sufficient to prime enzymatic amplification of said probe.

In an sixth aspect of the invention, there is provided a method of measuring gene expression in a sample derived from human adult liver, comprising:

contacting the single exon microarray in accordance with the second aspect of the invention, with a first collection of detectably labeled nucleic acids, said first collection of nucleic acids derived from mRNA of human adult liver; and then

measuring the label detectably bound to each probe of said microarray.

In a seventh aspect of the invention, there is provided a method of identifying exons in a eukaryotic genome, comprising:

algorithmically predicting at least one exon from genomic sequence of said eukaryote; and then

detecting specific hybridization of detectably labeled nucleic acids to a single exon probe,

wherein said detectably labeled nucleic acids are derived from mRNA from the adult liver of said eukaryote, said probe is a single exon probe having a fragment identical in sequence to, or complementary in sequence to, said predicted exon, said probe is included within a single exon microarray in accordance with the first aspect of the invention, and said fragment is selectively hybridizable at high stringency.

In a eighth aspect of the invention, there is provided a method of assigning exons to a single gene, comprising:

identifying a plurality of exons from genomic sequence in accordance with the seventh aspect of the invention; and then

measuring the expression of each of said exons in a plurality of tissues and/or cell types using hybridization to single exon microarrays having a probe with said exon,

wherein a common pattern of expression of said exons in said plurality of tissues and/or cell types indicates that the exons should be assigned to a single gene.

In an ninth aspect of the invention, there is provided a nucleic acid sequence as set out in any of SEQ ID NOs: 1 - 25,995wherein said sequence encodes a peptide.

In a tenth aspect of the invention, there is provided a peptide encoded by a sequence comprising a sequence as set out in any of SEQ ID NOs: 13,110 - 25,995, or a complementary sequence or coding portion thereof.

In a preferred embodiment, a peptide may be encoded by a sequence comprising a sequence set out in any of SEQ ID NOS.: 1 - 13,109.

In a further aspect, the invention provides peptides comprising an amino acid sequence translated from the DNA fragments, said amino acid sequences comprising SEQ

ID NOS.: 25,996 - 38,578.

Accordingly in a eleventh aspect of the invention there is provided a peptide comprising a sequence as set out in any of SEQ ID NOs: 25,996 - 38,578, or fragment thereof.

In another aspect, the invention provides means for displaying annotated sequence, and in particular, for displaying sequence annotated according to the methods and apparatus of the present invention. Further, such display can be used as a preferred graphical user interface for electronic search, query, and analysis of such annotated sequence.

## Detailed Description of the Invention

### Definitions

As used herein, the term "microarray" and phrase "nucleic acid microarray" refer to a substrate-bound collection of plural nucleic acids, hybridization to each of the plurality of bound nucleic acids being separately detectable. The substrate can be solid or porous, planar or non-planar, unitary or distributed.

As so defined, the term "microarray" and phrase "nucleic acid microarray" include all the devices so called in Schena (ed.), DNA Microarrays: A Practical Approach (Practical Approach Series), Oxford University Press (1999) (ISBN: 0199637768); *Nature Genet.* 21(1)(suppl):1 - 60 (1999); and Schena (ed.), Microarray Biochip: Tools and Technology, Eaton Publishing Company/BioTechniques Books Division (2000) (ISBN: 1881299376). As so defined, the term "microarray" and phrase "nucleic acid microarray" further include substrate-bound collections of plural nucleic acids in which the nucleic acids are distributably disposed on a plurality of beads, rather than on a unitary

planar substrate, as is described, *inter alia*, in Brenner *et al.*, *Proc. Natl. Acad. Sci. USA* 97(4):166501670 (2000); in such case, the term "microarray" and phrase "nucleic acid microarray" refer to the plurality of beads in aggregate.

As used herein with respect to a nucleic acid microarray, the term "probe" refers to the nucleic acid that is, or is intended to be, bound to the substrate; in such context, the term "target" thus refers to nucleic acid intended to be bound thereto by Watson-Crick complementarity. As used herein with respect to solution phase hybridization, the term "probe" refers to the nucleic acid of known sequence that is detectably labeled.

As used herein, the expression "probe comprising SEQ ID NO.", and variants thereof, intends a nucleic acid probe, at least a portion of which probe has either (i) the sequence directly as given in the referenced SEQ ID NO., or (ii) a sequence complementary to the sequence as given in the referenced SEQ ID NO., the choice as between sequence directly as given and complement thereof dictated by the requirement that the probe hybridize to mRNA.

As used herein, the term "open reading frame" and the equivalent acronym "ORF" refer to that portion of an exon that can be translated in its entirety into a sequence of contiguous amino acids i.e. a nucleic acid sequence that, in at least one reading frame, does not possess stop codons; the term does not require that the ORF encode the entirety of a natural protein.

As used herein, the term "amplicon" refers to a PCR product amplified from human genomic DNA, containing the predicted exon.

As used herein the term "exon" refers to the consensus prediction of the various exon and gene predicting algorithms i.e. a nucleic acid sequence bioinformatically predicted to encode a portion of a

natural protein.

As used herein, the term "peptide" refers to a sequence of amino acids. The sequences referred to as PEPTIDE SEQ ID NOS.: are the predicted peptide sequences that would be translated from one of the exons, or a portion thereof set out in exon SEQ ID NOS.:. The codons encoding the peptide are wholly contained within the exon.

As used herein, a "portions" of a defined nucleotide sequence or sequences can be and, preferably, are fragments unique to that sequence or to one or a combination of those sequences. A fragment unique to a nucleic acid molecule is one that is a signature for the larger nucleic acid molecule.

As used herein, the phrase "expression of a probe" and its linguistic variants means that the ORF present within the probe, or its complement, is present within a target mRNA.

As used herein, "stringent conditions" refers to parameters well known to those skilled in the art. When a nucleic acid molecule is said to be hybridisable to another of a given sequence under "stringent conditions" it is meant that it is homologous to the given sequence.

As used herein, the phrase "specific binding pair" intends a pair of molecules that bind to one another with high specificity. Binding pairs are said to exhibit specific binding when they exhibit avidity of at least $10^7$, preferably at least $10^8$, more preferably at least $10^9$ liters/mole. Nonlimiting examples of specific binding pairs are: antibody and antigen; biotin and avidin; and biotin and streptavidin.

As used herein with respect to the visual display of annotated genomic sequence, the term "rectangle" means any geometric shape that has at least a first and a second border, wherein the first and second borders each are capable of mapping uniquely to a point of another visual

object of the display.

As used herein, a "Mondrian" means a visual display in which a single genomic sequence is annotated with predicted and experimentally confirmed functional information.

## Brief Description of the Drawings

The present invention is further illustrated with reference to the following non-limiting figures and examples in which:

FIG. 1 illustrates a process for predicting functional regions from genomic sequence, confirming the functional activity of such regions experimentally, and associating and displaying the data so obtained in meaningful and useful relationship to the original sequence data;

FIG. 2 further elaborates that portion of the process schematized in FIG. 1 for predicting functional regions from genomic sequence;

FIG. 3 illustrates a Mondrian visual display;

FIG. 4 presents a Mondrian showing a hypothetical annotated genomic sequence;

FIG. 5 is a histogram showing the distribution of ORF length and PCR products as obtained, with ORF length shown in black and PCR product length shown in dotted lines;

FIG. 6 is a histogram showing the distribution, among exons predicted according to the methods described, of expression as measured using simultaneous two color hybridization to a genome-derived single exon microarray. The graph shows the number of sequence-verified products that were either not expressed ("0"), expressed in one or more but not all tested tissues ("1" - "9"), or expressed

in all tissues tested ("10");

FIG. 7 is a pictorial representation of the expression of verified sequences that showed expression with signal intensity greater than 3 in at least one tissue, with: FIG. 7A showing the expression as measured by microarray hybridization in each of the 10 measured tissues, and the expression as measured "bioinformatically" by query of EST, NR and SwissProt databases; with FIG. 7B showing the legend for display of physical expression (ratio) in FIG. 7A; and with FIG. 7C showing the legend for scoring EST hits as depicted in FIG. 7A;

FIG. 8 shows a comparison of normalized CY3 signal intensity for arrayed sequences that were identical to sequences in existing EST, NR and SwissProt databases or that were dissimilar (unknown), where black denotes the signal intensity for all sequence-verified products with a BLAST Expect ("E") value of greater than 1e-30 ($1 \times 10^{-30}$) ("unknown") and a dotted line denotes sequence-verified spots with a BLAST expect ("E") value of less than 1e-30 ($1 \times 10^{-30}$) ("known");

FIG. 9 presents a Mondrian of BAC AC008172 (bases 25,000 to 130,000), containing the carbamyl phosphate synthetase gene (AF154830.1); and

FIG. 10 is a Mondrian of BAC A049839.

Methods and Apparatus for Predicting, Confirming, Annotating, and Displaying Functional Regions From Genomic Sequence Data

FIG. 1 is a flow chart illustrating in broad outline a process for predicting functional regions from genomic sequence, confirming and characterizing the functional activity of such regions experimentally, and then associating and displaying the information so obtained

in meaningful and useful relationship to the original sequence data.

The initial input into process 10 of the present invention is drawn from one or more databases 100 containing genomic sequence data. Because genomic sequence is usually obtained from subgenomic fragments, the sequence data typically will be stored in a series of records corresponding to these subgenomic sequenced fragments. Some fragments will have been catenated to form larger contiguous sequences ("contigs"); others will not. A finite percentage of sequence data in the database will typically be erroneous, consisting *inter alia* of vector sequence, sequence created from aberrant cloning events, sequence of artificial polylinkers, and sequence that was erroneously read.

Each sequence record in database 100 will minimally contain as annotation a unique sequence identifier (accession number), and will typically be annotated further to identify the date of accession, species of origin, and depositor. Because database 100 can contain nongenomic sequence, each sequence will typically be annotated further to permit query for genomic sequence. Chromosomal origin, optionally with map location, can also be present. Data can be, and over time increasingly will be, further annotated with additional information, in part through use of the present invention, as described below. Annotation can be present within the data records, in information external to database 100 and linked to the records thereto, or through a combination of the two.

Databases useful as genomic sequence database 100 in the present invention include GenBank, and particularly include several divisions thereof, including the htgs(draft), NT (nucleotide, command line), and NR (nonredundant) divisions. GenBank is produced by the National Institutes of Health and is maintained by the

National Center for Biotechnology Information (NCBI). Databases of genomic sequence from species other than human, such as mouse, rat, Arabidopsis, *C. elegans*, *C. brigsii*, *Drosophila*, zebra fish, and other higher eukaryotic organisms will also prove useful as genomic sequence database 100.

Genomic sequence obtained by query of genomic sequence database 100 is then input into one or more processes 200 for identification of regions therein that are predicted to have a biological function as specified by the user. Such functions include, but are not limited to, encoding protein, regulating transcription, regulating message transport after transcription into mRNA, regulating message splicing after transcription into mRNA, of regulating message degradation after transcription into mRNA, and the like. Other functions include directing somatic recombination events, contributing to chromosomal stability or movement, contributing to allelic exclusion or X chromosome inactivation, and the like.

The particular genomic sequence to be input into process 200 will depend upon the function for which relevant sequence is to be identified as well as upon the approach chosen for such identification. Process step 200 can be iterated to identify different functions within a given genomic region. In such case, the input often will be different for the several iterations.

Sequences predicted to have the requisite function by process 200 are then input into process 300, where a subset of the input sequences suitable for experimental confirmation is identified. Experimental confirmation can involve physical and/or bioinformatic assay. Where the subsequent experimental assay is bioinformatic, rather than physical, there are fewer constraints on the sequences that can be tested, and in this latter case therefore process 300 can output the

entirety of the input sequence.

The subset of sequences output from process 300 is then used in process 400 for experimental verification and characterization of the function predicted in process 200, which experimental verification can, and often will, include both physical and bioinformatic assay.

Process 500 annotates the sequence data with the functional information obtained in the physical and/or bioinformatic assays of process 400. Such annotation can be done using any technique that usefully relates the functional information to the sequence, as, for example, by incorporating the functional data into the sequence data record itself, by linking records in a hierarchical or relational database, by linking to external databases, by a combination thereof, or by other means well known within the database arts. The data can even be submitted for incorporation into databases maintained by others, such as GenBank, which is maintained by NCBI.

As further noted in FIG. 1, additional annotation can be input into process 500 from external sources 600.

The annotated data is then displayed in process 800, either before, concomitantly with, or after optional storage 700 on nontransient media, such as magnetic disk, optical disc, magnetooptical disk, flash memory, or the like.

FIG. 1 shows that the experimental data output from process 400 can be used in each preceding step of process 10: e.g., facilitating identification of functional sequences in process 200, facilitating identification of an experimentally suitable subset thereof in process 300, and facilitating creation of physical and/or informational substrates for, and performance of subsequent assay, of functional sequences in process 400.

Information from each step can be passed directly to the succeeding process, or stored in permanent or

interim form prior to passage to the succeeding process. Often, data will be stored after each, or at least a plurality, of such process steps. Any or all process steps can be automated.

FIG. 2 further elaborates the prediction of functional sequence within genomic sequence according to process 200.

Genomic sequence database 100 is first queried 20 for genomic sequence.

The sequence required to be returned by query 20 will depend, in the first instance, upon the function to be identified.

For example, genomic sequences that function to encode protein can be identified *inter alia* using gene prediction approaches, comparative sequence analysis approaches, or combinations of the two. In gene prediction analysis, sequence from one genome is input into process 200 where at least one, preferably a plurality, of algorithmic methods are applied to identify putative coding regions. In comparative sequence analysis, by contrast, corresponding, *e.g.*, syntenic, sequence from a plurality of sources, typically a plurality of species, is input into process 200, where at least one, possibly a plurality, of algorithmic methods are applied to compare the sequences and identify regions of least variability.

The exact content of query 20 will also depend upon the database queried. For example, if the database contains both genomic and nongenomic sequence, perhaps derived from multiple species, and the function to be determined is protein coding regions in human genomic sequence, the query will accordingly require that the sequence returned be genomic and derived from humans.

Query 20 can also incorporate criteria that compel return of sequence that meets operative requirements of the subsequent analytical method. Alternatively, or in

addition, such operative criteria can be enforced in subsequent preprocess step 24.

For example, if the function sought to be identified is protein coding, query 20 can incorporate criteria that return from genomic sequence database 100 only those sequences present within contigs sufficiently long as to have obviated substantial fragmentation of any given exon among a plurality of separate sequence fragments.

Such criteria can, for example, consist of a required minimal individual genomic sequence fragment length, such as 10 kb, more typically 20 kb, 30 kb, 40kb, and preferably 50 kb or more, as well as an optional further or alternative requirement that sequence from any given clone, such as a bacterial artificial chromosome ("BAC"), be presented in no more than a finite maximal number of fragments, such as no more than 20 separate pieces, more typically no more than 15 fragments, even more typically no more than about 10 - 12 fragments.

Results using the present invention have shown that genomic sequence from bacterial artificial chromosomes (BACs) is sufficient for gene prediction analysis according to the present invention if the sequence is at least 50 kb in length, and if additionally the sequence from any given BAC is presented in fewer than 15, and preferably fewer than 10, fragments. Accordingly, query 20 can incorporate a requirement that data accessioned from BAC sequencing be in fewer than 15, preferably fewer than 10, fragments.

An additional criterion that can be incorporated into the query can be the date, or range of dates, of sequence accession. Although the process has been described above as if genomic sequence database 100 were static, it is of course understood that the genomic sequence databases need not be static, and indeed are typically updated on a frequent, even hourly, basis. Thus,

as further described in Examples 1 and 2, *infra*, it is possible to query the database for newly added sequence, either newly added after an absolute date, or newly added relative to a prior analysis performed using the methods and apparatus of the present invention. In this way, the process herein described can incorporate a dynamic, temporal component.

One utility of such temporal limitation is to identify, from newly accessioned genomic sequence, the presence of novel genes, particularly those not previously identified by EST sequencing (or other sequencing efforts that are similarly based upon gene expression). As further described in Example 1, such an approach has shown that newly accessioned human genomic sequence, when analyzed for sequences that function to encode protein, readily identifies genes that are novel over those in existing EST and other expression databases. This makes the methods of the present invention extremely powerful gene discovery tools. And as would be appreciated, such gene discovery can be performed using genomic sequence from species other than human.

If query 20 incorporates multiple criteria, such as above-described, the multiple criteria can be performed as a series of separate queries or as a single query, depending in part upon the query language, the complexity of the query, and other considerations well known in the database arts.

If query 20 returns no genomic sequence meeting the query criteria, the negative result can be reported by process 22, and process 200 (and indeed, entire process 10) ended 23, as shown. Alternatively, or in addition to report and termination of the initial inquiry, a new query 20 can be generated that takes into account the initial negative result.

When query 20 returns sequence meeting the query

criteria, the returned sequence is then passed to optional preprocessing 24, suitable and specific for the desired analytical approach and the particular analytical methods thereof to be used in process 25.

Preprocessing 24 can include processes suitable for many approaches and methods thereof, as well as processes specifically suited for the intended subsequent analysis.

Preprocessing 24 suitable for most approaches and methods will include elimination of sequence irrelevant to, or that would interfere with, the subsequent analysis. Such sequence includes repetitive sequence, such as Alu repeats and LINE elements, vector sequence, artificial sequence, such as artificial polylinkers, and the like. Such removal can readily be performed by identification and subsequent masking of the undesired sequence.

Identification can be effected by comparing the genomic sequence returned by query 20 with public or private databases containing known repetitive sequence, vector sequence, artificial sequence, and other artifactual sequence. Such comparison can readily be done using programs well known in the art, such as CROSS_MATCH, or by proprietary sequence comparison programs the engineering of which is well within the skill in the art.

Alternatively, or in addition, undesirable, including artifactual, sequence can be identified algorithmically without comparison to external databases and thereafter removed. For example, synthetic polylinker sequence can be identified by an algorithm that identifies a significantly higher than average density of known restriction sites. As another example, vector sequence can be identified by algorithms that identify nucleotide or codon usage at variance with that of the bulk of the genomic sequence.

Once identified, undesired sequence can be

removed. Removal can usefully be done by masking the undesired sequence as, for example, by converting the specific nucleotide references to one that is unrecognized by the subsequent bioinformatic algorithms, such as "X". Alternatively, but at present less preferred, the undesired sequence can be excised from the returned genomic sequence, leaving gaps.

Preprocessing 24 can further include selection from among duplicative sequences of that one sequence of highest quality. Higher quality can be measured as a lower percentage of, fewest number of, or least densely clustered occurrence of ambiguous nucleotides, defined as those nucleotides that are identified in the genomic sequence using symbols indicating ambiguity. Higher quality can also or alternatively be valued by presence in the longest contig.

Preprocessing 24 can, and often will, also include formatting of the data as specifically appropriate for passage to the analytical algorithms of process 25. Such formatting can and typically will include, *inter alia*, addition of a unique sequence identifier, either derived from the original accession number in genomic sequence database 100, or newly applied, and can further include additional annotation. Formatting can include conversion from one to another sequence listing standard, such as conversion to or from FASTA or the like, depending upon the input expected by the subsequent process.

Preprocessing, which can be optional depending upon the function desired to be identified and the informational requirements of the methods for effecting such identification, is followed by sequence processing 25, where sequences with the desired function are identified within the genomic sequence.

As mentioned above, such functions can include, but are not limited to, encoding protein, regulating

transcription, regulating message transport after transcription into mRNA, regulating message splicing after transcription, of regulating message degradation, and the like. Other functions include directing somatic recombination events, contributing to chromosomal stability or movement, contributing to allelic exclusion or X chromosome inactivation, or the like.

The methods of the present invention are particularly useful for gene discovery, that is, for identifying, from genomic sequence, regions that function to encode genes, and in a particularly useful embodiment, for identifying regions that function to encode genes not hitherto identified by expression-based or directed cloning and sequencing. In conjunction with verification using the novel single exon microarrays of the present invention, as further described below, the methods herein described become powerful gene discovery tools.

Accordingly, in a preferred embodiment of the present invention, process 25 is used to identify putative coding regions. Two preferred approaches in process 25 for identifying sequence that encodes putative genes are gene prediction and comparative sequence analysis.

Gene prediction can be performed using any of a number of algorithmic methods, embodied in one or more software programs, that identify open reading frames (ORFs) using a variety of heuristics, such as GRAIL, DICTION, and GENEFINDER. Comparative sequence analysis similarly can be performed using any of a variety of known programs that identify regions with lower sequence variability.

As further described in Example 1, below, gene finding software programs yield a range of results. For the newly accessioned human genomic sequence input in Example 1, for example, GRAIL identified the greatest percentage of genomic sequence as putative coding region, 2% of the data analyzed; GENEFINDER was second, calling 1%;

and DICTION yielded the least putative coding region, with 0.8% of genomic sequence called as coding region.

Increased reliability can be obtained when consensus is required among several such methods. Although discussed herein particularly with respect to exon calling, consensus among methods will in general increase reliability of predicting other functions as well.

Thus, as indicated by query 26, sequence processing 25, optionally with preprocessing 24, can be repeated with a different method, with consensus among such iterations determined and reported in process 27.

Process 27 compares the several outputs for a given input genomic sequence and identifies consensus among the separately reported results. The consensus itself, as well as the sequence meeting that consensus, is then stored in process 29a, displayed in process 29b, and/or output to process 300 for subsequent identification of a subset thereof suitable for assay.

Multiple levels of consensus can be calculated and reported by process 27. For example, as further described in Example 1, *infra*, process 27 can report consensus as between all specific pairs of methods of gene prediction, as consensus among any one or more of the pairs of methods of gene prediction, or as among all of the gene prediction algorithms used. Thus, in Example 1, process 27 reported that GRAIL and GENEFINDER programs agreed on 0.7% of genomic sequence, that GRAIL and DICTION agreed on 0.5% of genomic sequence, and that the three programs together agreed on 0.25% of the data analyzed. Put another way, 0.25% of the genomic sequence was identified by all three of the programs as containing putative coding region.

Furthermore, consensus can be required among different approaches to identifying a chosen function.

For example, if the function desired to be identified is coding of protein sequence, and a first used

approach to exon calling is gene prediction, the process can be repeated on the same input sequence, or subset thereof, with another approach, such as comparative sequence analysis. In such a case, where comparative sequence analysis follows gene prediction, the comparison can be performed not only on genomic nucleic acid sequence, but additionally or alternatively can be performed on the predicted amino acid sequence translated from the ORFs prior identified by the gene prediction approach.

Although shown as an iterative process, the multiple analyses required to achieve consensus can be done in series, in parallel, or some combination thereof.

Predicted functional sequence, optionally representing a consensus among a plurality of methods and approaches for determination thereof, is passed to process 300 for identification of a subset thereof for functional assay.

In the preferred embodiment of the methods of the present invention, wherein the function sought to be identified is protein coding, process 300 is used to identify a subset thereof suitable for experimental verification by physical and/or bioinformatic approaches.

For example, putative ORFs identified in process 200 can be classified, or binned, bioinformatically into putative genes. This binning can be based *inter alia* upon consideration of the average number of exons/gene in the species chosen for analysis, upon density of exons that have been called on the genomic sequence, and other empirical rules. Thereafter, one or more among the gene-specific ORFs can be chosen for subsequent use in gene expression assay.

Where such subsequent gene expression assay uses amplified nucleic acid, considerations such as desired amplicon length, primer synthesis requirements, putative exon length, sequence GC content, existence of possible

secondary structure, and the like can be used to identify and select those ORFs that appear most likely successfully to amplify. Where subsequent gene expression assay relies upon nucleic acid hybridization, whether or not using amplified product, further considerations involving hybridization stringency can be applied to identify that subset of sequences that will most readily permit sequence-specific discrimination at a chosen hybridization and wash stringency. One particular such consideration is avoidance of putative exons that span repetitive sequence; such sequence can hybridize spuriously to nonspecific message, reducing specific signal in the hybridization.

For bioinformatic assay, there are fewer constraints on the sequences that can be tested experimentally, and in this latter case therefore process 300 can output the entirety of the input sequence.

The subset of sequences identified by process 300 as suitable for use in assay is then used in process 400 to create the physical and/or informational substrate for experimental verification of the predictions made in process 200, and thereafter to assay those substrates.

As mentioned, the methods of the present invention are particularly useful for identifying potential coding regions within genomic sequence. In a preferred embodiment of process 400, therefore, the expression of the sequences predicted to encode protein is verified. The combination of the predictive and experimental methods provides a powerful gene discovery engine.

Thus, in another aspect, the present invention provides methods and apparatus for verifying the expression of putative genes identified within genomic sequence. In particular, the invention provides a novel method of verifying gene expression in which expression of predicted ORFs is measured and confirmed using a novel type of nucleic acid microarray, the genome-derived single exon

nucleic acid microarrays of the present invention.

Putative ORFs as predicted by a consensus of gene calling, particularly gene prediction, algorithms in process 200, and as further identified as suitable by process 300, are amplified from genomic DNA using the polymerase chain reaction (PCR). Although PCR is conveniently used, other amplification approaches can also be used.

Amplification schemes can be designed to capture the entirety of each predicted ORF in an amplicon with minimal additional (that is, intronic or intergenic) sequence. Because ORFs predicted from human genomic sequence using the methods of the present invention differ in length, such an approach results in amplicons of varying length.

However, most predicted ORFs are shorter than 500 bp in length, and although amplicons of at least about 100 or 200 base pairs can be immobilized as probes on nucleic acid microarrays, early experimental results using the methods of the present invention have suggested that longer amplicons, at least about 400 or 500 base pairs, are more effective. Furthermore, certain advantages derive from application to the microarray of amplicons of defined size.

Therefore, amplification schemes can alternatively, and preferably, be designed to amplify regions of defined size, preferably at least about 300, 400 or 500 bp, centered about each predicted ORF. Such an approach results in a population of amplicons of limited size diversity, but that typically contain intronic and/or intergenic nucleic acid in addition to putative ORF.

Conversely, somewhat fewer than 10% of ORFs predicted from human genomic sequence according to the methods of the present invention exceed 500 bp in length. Portions of such extended ORFs, preferably at least about 300,400 or 500 bp in length, can be amplified. However, it

has been discovered that the percentage success at amplifying pieces of such ORFs is low, and that such putative exons are more effectively amplified when larger fragments, at least about 1000 or 1500 bp, and even as large as 2000 bp are amplified.

The putative ORFs selected in process 300 are thus input into one or more primer design programs, such as PRIMER3 (available online for use at http://www-genome.wi.mit.edu/cgi-bin/primer/ ), with a goal of amplifying at least about 500 base pairs of genomic sequence centered within or about ORFs predicted to be no more than about 500 bp, or at least about 1000 - 1500 bp of genomic sequence for ORFs predicted to exceed 500 bp in length, and the primers synthesized by standard techniques. Primers with the requisite sequences can be purchased commercially or synthesized by standard techniques.

Conveniently, a first predetermined sequence can be added commonly to the ORF-specific 5' primer and a second, typically different, predetermined sequence commonly added to each 3' ORF-unique primer. This serves to immortalize the amplicon, that is, serves to permit further amplification of any amplicon using a single set of primers complementary respectively to the common 5' and common 3' sequence elements. The presence of these "universal" priming sequences further facilitates later sequence verification, providing a sequence common to all amplicons at which to prime sequencing reactions. The common 5' and 3' sequences further serve to add a cloning site should any of the ORFs warrant further study.

Such predetermined sequence is usefully at least about 10, 12 or 15 nt in length, and usually does not exceed about 25 nt in length. The "universal" priming sequences used in the examples presented *infra* were each 16 nt long.

The genomic DNA to be used as substrate for

amplification will come from the eukaryotic species from which the genomic sequence data had originally been obtained, or a closely related species, and can conveniently be prepared by well known techniques from somatic or germline tissue or cultured cells of the organism. See, *e.g.*, Short Protocols in Molecular Biology : A Compendium of Methods from Current Protocols in Molecular Biology, Ausubel et al. (eds.), 4th edition (April 1999), John Wiley & Sons (ISBN: 047132938X) and Maniatis *et al.*, Molecular Cloning : A Laboratory Manual, 2nd edition (December 1989), Cold Spring Harbor Laboratory Press (ISBN: 0879693096). Many such prepared genomic DNAs are available commercially, with the human genomic DNAs additionally having certification of donor informed consent.

Although the intronic and intergenic material flanking putative coding regions in the amplicons could potentially interfere with hybridizations during microarray experiments, we have found, surprisingly, that differential expression ratios are not significantly affected. Rather, the predominant effect of exon size is to alter the absolute signal intensity, rather than its ratio. Equally surprising, the art had suggested that single exon probes would not provide sufficient signal intensity for high stringency hybridization analyses; we find that such probes not only provide adequate signal, but have substantial advantages, as herein described.

After partial purification, as by size exclusion spin column, with or without confirmation as to amplicon quality as by gel electrophoresis, each amplicon (single exon probe) is disposed in an array upon a support substrate.

Methods for creating microarrays by deposition and fixation of nucleic acids onto support substrates are well known in the art (Reviewed by Schena et al., see

above).

Typically, the support substrate will be glass, although other materials, such as amorphous or crystalline silicon or plastics. Such plastics include polymethylacrylic, polyethylene, polypropylene, polyacrylate, polymethylmethacrylate, polyvinylchloride, polytetrafluoroethylene, polystyrene, polycarbonate, polyacetal, polysulfone, celluloseacetate, cellulosenitrate, nitrocellulose, or mixtures thereof, can also be used. Typically, the support will be rectangular, although other shapes, particularly circular disks and even spheres, present certain advantages. Particularly advantageous alternatives to glass slides as support substrates for array of nucleic acids are optical discs, as described in WO 98/12559.

The amplified nucleic acids can be attached covalently to a surface of the support substrate or, more typically, applied to a derivatized surface in a chaotropic agent that facilitates denaturation and adherence by presumed noncovalent interactions, or some combination thereof.

Robotic spotting devices useful for arraying nucleic acids on support substrates can be constructed using public domain specifications (The MGuide, version 2.0, http://cmgm.stanford.edu/pbrown/mguide/index.html), or can conveniently be purchased from commercial sources (MicroArray GenII Spotter and MicroArray GenIII Spotter, Molecular Dynamics, Inc., Sunnyvale, CA). Spotting can also be effected by printing methods, including those using ink jet technology.

As is well known in the art, microarrays typically also contain immobilized control nucleic acids. For controls useful in providing measurements of background signal for the genome-derived single exon microarrays of the present invention, a plurality of *E. coli* genes can

readily be used. As further described in Example 1, 16 or 32 *E. coli* genes suffice to provide a robust measure of background noise in such microarrays.

As is well known in the art, the amplified product disposed in arrays on a support substrate to create a nucleic acid microarray can consist entirely of natural nucleotides linked by phosphodiester bonds, or alternatively can include either nonnative nucleotides, alternative internucleotide linkages, or both, so long as complementary binding can be obtained in the hybridization. If enzymatic amplification is used to produce the immobilized probes, the amplifying enzyme will impose certain further constraints upon the types of nucleic acid analogs that can be generated.

Although particularly described herein as using high density microarrays constructed on planar substrates, the methods of the present invention for confirming the expression of ORFs predicted from genomic sequence can use any of the known types of microarrays, as herein defined, including lower density planar arrays, and microarrays on nonplanar, nonunitary, distributed substrates.

For example, gene expression can be confirmed using hybridization to lower density arrays, such as those constructed on membranes, such as nitrocellulose, nylon, and positively-charged derivatized nylon membranes. Further, gene expression can also be confirmed using nonplanar, bead-based microarrays such as are described in Brenner *et al.*, *Proc. Natl. Acad. Sci. USA* 97(4):166501670 (2000); U.S. Patent No. 6,057,107; and U.S. Patent No. 5,736,330. In theory, a packed collection of such beads provides in aggregate a higher density of nucleic acid probe than can be achieved with spotting or lithography techniques on a single planar substrate.

Planar microarrays on solid substrates, however, provide certain useful advantages, including high

throughput and compatibility with existing readers. For example, each standard microscope slide can include at least 1000, typically at least 2000, preferably 5000 and upto 10,000 - 50,000 or more nucleic acid probes of discrete sequence. The number of sequences deposited will depend on their required application.

Each putative gene can be represented in the array by a single predicted ORF. Alternatively, genes can be represented by more than one predicted ORF. For purposes of measuring differential splicing, more than one predicted ORF will be provided for a putative gene. And as is well known in the art, each probe of defined sequence, representing a single predicted ORF, can be deposited in a plurality of locations on a single microarray to provide redundancy of signal.

The genome-derived single exon microarrays described above differ in several fundamental and advantageous ways from microarrays presently used in the gene expression art, including (1) those created by deposition of mRNA-derived nucleic acids, (2) those created by *in situ* synthesis of oligonucleotide probes, and (3) those constructed from yeast genomic DNA.

Most nucleic acid microarrays that are in use for study of eukaryotic gene expression have as immobilized probes nucleic acids that are derived – either directly or indirectly – from expressed message. As discussed above, it is common, for example, for such microarrays to be derived from cDNA/EST libraries, either from those previously described in the literature, see Lennon *et al.*, or from the *de novo* construction of "problem specific" libraries targeted at a particular biological question, R.S. Thomas *et al.*, *Cancer Res.* (in press). Such microarrays are herein collectively denominated "EST microarrays".

Such EST microarrays by definition can measure

expression only of those genes found in EST libraries, shown herein to represent only a fraction of expressed genes. Furthermore, such libraries – and thus microarrays based thereupon – are biased by the tissue or cell type of message origin, by the expression levels of the respective genes within the tissues, and by the ability of the message successfully to have been reverse-transcribed and cloned.

Thus, as further discussed in Example 1, the methods of the present invention enable sequences that do not appear in EST or other expression databases to be determined - subsequently arrayed for expression measurements could not, therefore, have been represented as probes on an EST microarray. And as further demonstrated in the examples, *infra*, the remaining population of genes identified from genomic sequence by the methods of the present invention – that is, the one third of sequences that had previously been accessioned in EST or other expression databases – are biased toward genes with higher expression levels.

Representation of a message in an EST and/or cDNA library depends upon the successful reverse transcription, optionally but typically with subsequent successful cloning, of the message. This introduces substantial bias into the population of probes available for arraying in EST microarrays.

In contrast, neither reverse transcription nor cloning is required to produce the probes arrayed on the genome-derived single exon microarrays of the present invention. And although the ultimate deposition of a probe on the genome-derived single exon microarray of the present invention depends upon a successful amplification from genomic material, *a priori* knowledge of the sequence of the desired amplicon affords greater opportunity to recover any given probe sequence recalcitrant to amplification than is afforded by the requirement for successful reverse

transcription and cloning of unknown message in EST approaches.

Thus, the genome-derived single exon microarrays of the present invention present a far greater diversity of probes for measuring gene expression, with far less bias, than do EST microarrays presently used in the art.

As a further consequence of their ultimate origin from expressed message, the probes in EST microarrays often contain poly-A (or complementary poly-T) stretches derived from the poly-A tail of mature mRNA. These homopolymeric stretches contribute to cross-hybridization, that is, to a spurious signal occasioned by hybridization to the homopolymeric tail of a labeled cDNA that lacks sequence homology to the gene-specific portion of the probe.

In contrast, the probes arrayed in the genome-derived single exon microarrays of the present invention lack homopolymeric stretches derived from message polyadenylation, and thus can provide more specific signal. Typically, at least about 50, 60 or 75% of the probes on the genome-derived single exon microarrays of the present invention lack homopolymeric regions consisting of A or T, where a homopolymeric region is defined for purposes herein as stretches of 25 or more, typically 30 or more, identical nucleotides.

A further distinction, which also affects the specificity of hybridization, is occasioned by the typical derivation of EST microarray probes from cloned material. Because much of the probe material disposed as probes on EST microarrays is excised or amplified from plasmid, phage, or phagemid vectors, EST microarrays typically include a fair amount of vector sequence, more so when the probes are amplified, rather than excised, from the vector.

In contrast, the vast majority of probes in the genome-derived single exon microarrays of the present invention contain no prokaryotic or bacteriophage vector

sequence, having been amplified directly or indirectly from genomic DNA. Typically, therefore, at least about 50, 60, 70 or 80% or more of individual exon-including probes disposed on a genome-derived single exon microarray of the present invention lack vector sequence, and particularly lack sequences drawn from plasmids and bacteriophage. Preferably, at least about 85, 90 or more than 90% of exon-including probes in the genome-derived single exon microarray of the present invention lack vector sequence. With attention to removal of vector sequences through preprocessing 24, percentages of vector-free exon-including probes can be as high as 95 - 99%. The substantial absence of vector sequence from the genome-derived single exon microarrays of the present invention results in greater specificity during hybridization, since spurious cross-hybridization to a probe vector sequence is reduced.

As a further consequence of excision or amplification of probes from vectors in construction of EST microarrays, the probes arrayed thereon often contain artificial sequence, derived from vector polylinker multiple cloning sites, at both 5' and 3' ends. The probes disposed upon the genome-derived single exon microarrays need have no such artificial sequence appended thereto.

As mentioned above, however, the ORF-specific primers used to amplify putative ORFs can include artificial sequences, typically 5' to the ORF-specific primer sequence, useful for "universal" (that is, independent of ORF sequence) priming of subsequent amplification or sequencing reactions. When such "universal" 5' and/or 3' priming sequences are appended to the amplification primers, the probes disposed upon the genome-derived single exon microarray will include artificial sequence similar to that found in EST microarrays. However, the genome-derived single exon microarray of the present invention can be made without

such sequences, and if so constructed, presents an even smaller amount of nonspecific sequence that would contribute to nonspecific hybridization.

Yet another consequence of typical use of cloned material as probes in EST microarrays is that such microarrays contain probes that result from cloning artifacts, such as chimeric molecules containing coding region of two separate genes. Derived from genomic material, typically not thereafter cloned, the probes of the genome-derived single exon microarrays of the present invention lack such cloning artifacts, and thus provide greater specificity of signal in gene expression measurements.

A further consequence of the cloned origin of probes on many EST microarrays is that the individual probes often have disparate sizes, which can cause the optimal hybridization stringency to vary among probes on a single microarray. In contrast, as discussed above, the probes arrayed on the genome-derived single exon microarrays of the present invention can readily be designed to have a narrow distribution in sizes, with the range of probe sizes no greater than about 10% of the average size, typically no greater than about 5% of the average probe size.

Because of their origin from fully- or partially-spliced message, probes disposed upon EST arrays will often include multiple exons. The percentage of such exon-spanning probes in an EST microarray can be calculated, on average, based upon the predicted number of exons/gene for the given species and the average length of the immobilized probes. For human genes, the near-complete sequence of human chromosome 22, Dunham *et al.*, *Nature* 402(6761):489-95 (1999), predicts that human genes average 5.5 exons/gene. Even with probes of 200 - 500 bp, the vast majority of human EST microarray probes include more than one exon.

In contrast, by virtue of their origin from algorithmically identified ORFs in genomic sequence, the probes in the genome-derived single exon microarrays of the present invention can consist of individual exons. Thus, in contrast to EST microarrays, at least about 50, 60, 70, 75, 80, 85, 95 or 99% of probes deposited in the genome-derived microarray of the present invention consist of, or include, no more than one predicted ORF.

This provides the ability, not readily achieved using EST microarrays, to use the genome-derived single exon microarrays of the present invention to measure tissue-specific expression of individual exons, which in turn allows differential splicing events to be detected and characterized, and in particular, allows the correlation of differential splicing to tissue-specific expression patterns.

Furthermore, the exons that are represented in EST microarrays are often biased toward the 3' or 5' end of their respective genes, since sequencing strategies used for EST identification are so biased. In contrast, no such 3' or 5' bias necessarily inheres in the selection of exons for disposition on the genome-derived single exon microarrays of the present invention.

Conversely, the probes provided on the genome-derived single exon microarrays of the present invention typically, but need not necessarily, include intronic and/or intergenic sequence that is absent from EST microarrays, which are derived from mature mRNA. Typically, at least about 50, 60, 70, 80 or 90% of the exon-including probes on the genome-derived single exon microarrays of the present invention include sequence drawn from noncoding regions. As discussed above, the additional presence of noncoding region does not significantly interfere with measurement of gene expression, and provides the additional opportunity to assay prespliced RNA, and

thus measure such phenomena such as nuclear export control.

The genome-derived single exon microarrays of the present invention are also quite different from *in situ* synthesis microarrays, where probe size is severely constrained by inadequacies in the photolithographic synthesis process.

Typically, probes arrayed on *in situ* synthesis microarrays are limited to a maximum of about 25 bp. As a well known consequence, hybridization to such chips must be performed at low stringency. In order, therefore, to achieve unambiguous sequence-specific hybridization results, the *in situ* synthesis microarray requires substantial redundancy, with concomitant programmed arraying for each probe of probe analogues with altered (*i.e.*, mismatched) sequence.

In contrast, the longer probe length of the genome-derived single exon microarrays of the present invention allows much higher stringency hybridization and wash. Typically, therefore, exon-including probes on the genome-derived single exon microarrays of the present invention average at least about 100, 200, 300, 400 or 500 bp in length. By obviating the need for substantial probe redundancy, this approach permits a higher density of probes for discrete exons or genes to be arrayed on the microarrays of the present invention than can be achieved for *in situ* synthesis microarrays.

A further distinction is that the probes in *in situ* synthesis microarrays typically are covalently linked to the substrate surface. In contrast, the probes disposed on the genome-derived microarray of the present invention typically are, but need not necessarily be, bound noncovalently to the substrate.

Furthermore, the short probe size on *in situ* microarrays causes large percentage differences in the melting temperature of probes hybridized to their

complementary target sequence, and thus causes large percentage differences in the theoretically optimum stringency across the array as a whole.

In contrast, the larger probe size in the microarrays of the present invention create lower percentage differences in melting temperature across the range of arrayed probes.

A further significant advantage of the microarrays of the present invention over *in situ* synthesized arrays is that the quality of each individual probe can be confirmed before deposition. In contrast, the quality of probes cannot be assessed on a probe-by-probe basis for the *in situ* synthesized microarrays presently being used.

The genome-derived single exon microarrays of the present invention are also distinguished over, and present substantial benefits over, the genome-derived microarrays from lower eukaryotes such as yeast. Lashkari *et al.*, *Proc. Natl. Acad. Sci. USA* 94:13057-13062 (1997).

Only about 220 - 250 of the 6100 or so nuclear genes in *Saccharomyces cerevisiae* — that is, only about 4 - 5% — have standard, spliceosomal, introns, Lopez *et al.*, *Nucl. Acids Res.* 28:85-86 (2000); Spingola *et al.*, *RNA* 5(2):221-34 (1999). Furthermore, the entire yeast genome has already been sequenced. These two facts permit the ready amplification and disposition of single-ORF amplicons on such microarray without the requirement for antecedent use of gene prediction and/or comparative sequence analyses.

Thus, a significant aspect of the present invention is the ability to identify and to confirm expression of predicted coding regions in genomic sequence drawn from eukaryotic organisms that have a higher percentage of genes having introns than do yeast such as *Saccharomyces cerevisiae*, particularly in genomic sequence

drawn from eukaryotes in which at least about 10, 20 or 50% of protein-encoding genes have introns. In preferred embodiments, the methods and apparatus of the present invention are used to identify and confirm expression of novel genes from genomic sequence of eukaryotes in which the average number of introns per gene is at least about one, two or three or more.

After the physical substrate is prepared, experimental verification of predicted function is performed.

In a preferred embodiment of the present invention, where the function sought to be identified in genomic sequence is protein coding, experimental verification is performed by measuring expression of the putative ORFs, typically through nucleic acid hybridization experiments, and in particularly preferred embodiments, through hybridization to genome-derived single exon microarrays prepared as above- described.

Expression is conveniently measured and expressed for each probe in the microarray as a ratio of the expression measured concurrently in a plurality of mRNA sources, according to techniques well known in the microarray art, *Reviewed in* Schena et al., and as further described in Example 2, below. The mRNA source for the reference against which specific expression is measured can be drawn from a homogeneous mRNA source, such as a single cultured cell-type, or alternatively can be heterogeneous, as from a pool of mRNA derived from multiple tissues and/or cell types, as further described in Example 2, *infra*.

mRNA can be prepared by standard techniques, see Ausubel et al. and Maniatis *et al.*, or purchased commercially. The mRNA is then typically reverse-transcribed in the presence of labeled nucleotides: the index source (that in which expression is desired to be measured) is reverse transcribed in the presence of

nucleotides labeled with a first label, typically a fluorophore (fluorochrome; fluor; fluorescent dye); the reference source is reverse transcribed in the presence of a second label, typically a fluorophore, typically fluorometrically-distinguishable from the first label. As further described in Example 2, *infra*, Cy3 and Cy5 dyes prove particularly useful in these methods. After partial purification of the index and reference targets, hybridization to the probe array is conducted according to standard techniques, typically under a coverslip.

After wash, microarrays are conveniently scanned using a commercial microarray scanning device, such as a Gen3 Scanner (Molecular Dynamics, Sunnyvale, CA). Data on expression is then passed, with or without interim storage, to process 500, where the results for each probe are related to the original sequence.

Often, hybridization of target material to the genome-derived single exon microarray will identify certain of the probes thereon as of particular interest. Thus, it is often desirable that the user be able readily to obtain sufficient quantities of an individual probe, either for subsequent arrayed deposition upon an additional support substrate, often as part of a microarray having a plurality of probes so identified, or alternatively or additionally as a solitary solid-phase or solution-phase probe, for further use.

Thus, in another aspect, the present invention provides compositions and kits for the ready production of nucleic acids identical in sequence to, or substantially identical in sequence to, probes on the genome-derived single exon microarrays of the present invention.

In this aspect, a small quantity of each probe is disposed, typically without attachment to substrate, in a spatially-addressable ordered set, typically one per well of a microtiter dish. Although a 96 well microtiter plate

can be used, greater efficiency is obtained using higher density arrays, such as are provided by microtiter plates having 384, 864, 1536, 3456, 6144, or 9600 wells, and although microtiter plates having physical depressions (wells) are conveniently used, any device that permits addressable withdrawal of reagent from fluidly-noncommunicating areas can be used.

In this aspect of the invention, therefore, a fluidly noncommunicating addressable ordered set of individual probes, corresponding to those on a genome-derived single exon microarray, is provided, with each probe in sufficient quantity to permit amplification, such as by PCR. As earlier mentioned, the ORF-specific 5' primers used for genomic amplification can have a first common sequence added thereto, and the ORF-specific 3' primers used for genomic amplification can have a second, different, common sequence added thereto, thus permitting, in this preferred embodiment, the use of a single set of 5' and 3' primers to amplify any one of the probes from the amplifiable ordered set.

Each discrete amplifiable probe can also be packaged with amplification primers, solutes, buffers, *etc.*, and can be provided in dry (*e.g.*, lyophilized) form or wet, in the latter case typically with addition of agents that retard evaporation.

In another aspect of the present invention, a genome-derived single-exon microarray is packaged together with such an ordered set of amplifiable probes corresponding to the probes, or one or more subsets of probes, thereon. In alternative embodiments, the ordered set of amplifiable probes is packaged separately from the genome-derived single exon microarray.

In some embodiments, the microarray and/or ordered probe set are further packaged with recordable media that provide probe identification and addressing

information, and that can additionally contain annotation information, such as gene expression data. Such recordable media can be packaged with the microarray, with the ordered probe set, or with both.

If the microarray is constructed on a substrate that incorporates recordable media, such as is described in international patent application no. WO 98/12559, then separate packaging of the genome-derived single exon microarray and the bioinformatic information is not required.

The amount of amplifiable probe material should be sufficient to permit at least one amplification sufficient for subsequent hybridization assay.

Although the use of high density genome-derived microarrays on solid planar substrates is presently a preferred approach for the physical confirmation and characterization of the expression of sequences predicted to encode protein, other types of microarrays (as herein defined) can also be used.

Furthermore, as earlier mentioned, experimental verification of the function predicted from genomic sequence in process 200 can be bioinformatic, rather than, or additional to, physical verification.

For example, where the function desired to be identified is protein coding, the predicted ORFs can be compared bioinformatically to sequences known or suspected of being expressed.

Thus, the sequences output from process 300 (or process 200), can be used to query expression databases, such as EST databases, SNP ("single nucleotide polymorphism") databases, known cDNA and mRNA sequences, SAGE ("serial analysis of gene expression") databases, and more generalized sequence databases that allow query for expressed sequences. Such query can be done by any sequence query algorithm, such as BLAST ("basic local

alignment search tool"). The results of such query – including information on identical sequences and information on nonidentical sequences that have diffuse or focal regions of sequence homology to the query sequence – can then be passed directly to process 500, or used to inform analyses subsequently undertaken in process 200, process 300, or process 400.

Experimental data, whether obtained by physical or bioinformatic assay in process 400, is passed to process 500 where it is usefully related to the sequence data itself, a process colloquially termed "annotation". Such annotation can be done using any technique that usefully relates the functional information to the sequence, as, for example, by incorporating the functional data into the record itself, by linking records in a hierarchical or relational database, by linking to external databases, or by a combination thereof. Such database techniques are well within the skill in the art.

The annotated sequence data can be stored locally, uploaded to genomic sequence database 100, and/or displayed 800.

The methods and apparatus of the present invention rapidly produce functional information from genomic sequence. Coupled with the escalating pace at which sequence now accumulates, the rapid pace of sequence annotation produces a need for methods of displaying the information in meaningful ways.

FIG. 3 shows visual display 80 presenting a single genomic sequence annotated according to the present invention. Because of its nominal resemblance to artistic works of Piet Mondrian, visual display 80 is alternatively described herein as a "Mondrian".

Each of the visual elements of display 80 is aligned with respect to the genomic sequence being annotated (hereinafter, the "annotated sequence"). Given

the number of nucleotides typically represented in an annotated sequence, representation of individual nucleotides would rarely be readable in hard copy output of display 80. Typically, therefore, the annotated sequence is schematized as rectangle 89, extending from the left border of display 80 to its right border. By convention herein, the left border of rectangle 89 represents the first nucleotide of the sequence and the right border of rectangle 89 represents the last nucleotide of the sequence.

As further discussed below, however, the Mondrian visual display of annotated sequence can serve as a convenient graphical user interface for computerized representation, analysis, and query of information stored electronically. For such use, the individual nucleotides can conveniently be linked to the X axis coordinate of rectangle 89. This permits the annotated sequence at any point within rectangle 89 readily to be viewed, either automatically – for example, by time-delayed appearance of a small overlaid window upon movement of a cursor or other pointer over rectangle 89 – or through user intervention, as by clicking a mouse or other pointing device at a point in rectangle 89.

Visual display 80 is generated after user specification of the genomic sequence to be displayed. Such specification can consist of or include an accession number for a single clone (*e.g.*, a single BAC accessioned into GenBank), wherein the starting and stopping nucleotides are thus absolutely identified, or alternatively can consist of or include an anchor or fulcrum point about which a chosen range of sequence is anchored, thus providing relative endpoints for the sequence to be displayed. For example, the user can anchor such a range about a given chromosomal map location, gene name, or even a sequence returned by query for similarity

or identity to an input query sequence. When visual display 80 is used as a graphical user interface to computerized data, additional control over the first and last displayed nucleotide will typically be dynamically selectable, as by use of standard zooming and/or selection tools.

Field 81 of visual display 80 is used to present the output from process 200, that is, to present the bioinformatic prediction of those sequences having the desired function within the genomic sequence. Functional sequences are typically indicated by at least one rectangle 83 (83a, 83b, 83c), the left and right borders of which respectively indicate, by their X-axis coordinates, the starting and ending nucleotides of the region predicted to have function.

Where a single bioinformatic method or approach identifies a plurality of regions having the desired function, a plurality of rectangles 83 is disposed horizontally in field 81. Where multiple methods and/or approaches are used to identify function, each such method and/or approach can be represented by its own series of horizontally disposed rectangles 83, each such horizontally disposed series of rectangles offset vertically from those representing the results of the other methods and approaches.

Thus, rectangles 83a in FIG. 3 represent the functional predictions of a first method of a first approach for predicting function, rectangles 83b represent the functional predictions of a second method and/or second approach for predicting that function, and rectangles 83c represent the predictions of a third method and/or approach.

Where the function desired to be identified is protein coding, field 81 is used to present the bioinformatic prediction of sequences encoding protein.

For example, rectangles 83a can represent the results from GRAIL or GRAIL II, rectangles 83b can represent the results from GENEFINDER, and rectangles 83c can represent the results from DICTION.

Optionally, and preferably, rectangles 83 collectively representing predictions of a single method and/or approach are identically colored and/or textured, and are distinguishable from the color and/or texture used for a different method and/or approach.

Alternatively, or in addition, the color, hue, density, or texture of rectangles 83 can be used further to report a measure of the bioinformatic reliability of the prediction. For example, many gene prediction programs will report a measure of the reliability of prediction. Thus, increasing degrees of such reliability can be indicated, *e.g.*, by increasing density of shading. Where display 80 is used as a graphical user interface, such measures of reliability, and indeed all other results output by the program, can additionally or alternatively be made accessible through linkage from individual rectangles 83, as by time-delayed window ("tool tip" window), or by pointer (*e.g.*, mouse)-activated link.

As earlier described, increased predictive reliability can be achieved by requiring consensus among methods and/or approaches to determining function. Thus, field 81 can include a horizontal series of rectangles 83 that indicate one or more degrees of consensus in predictions of function.

Although FIG. 3 shows three series of horizontally disposed rectangles in field 81, display 80 can include as few as one such series of rectangles and as many as can discriminably be displayed, depending upon the number of methods and/or approaches used to predict a given function.

Furthermore, field 81 can be used to show

predictions of a plurality of different functions. However, the increased visual complexity occasioned by such display makes more useful the ability of the user to select a single function for display. When display 80 is used as a graphical user interface for computer query and analysis, such function can usefully be indicated and user-selectable, as by a series of graphical buttons or tabs (not shown in FIG. 3).

Rectangle 89 is shown in FIG. 3 as including interposed rectangle 84. Rectangle 84 represents the portion of annotated sequence for which predicted functional information has been assayed physically, with the starting and ending nucleotides of the assayed material indicated by the X axis coordinates of the left and right borders of rectangle 84. Rectangle 85, with optional inclusive circles 86 (86a, 86b, and 86c) displays the results of such physical assay.

Although a single rectangle 84 is shown in FIG. 3, physical assay is not limited to just one region of annotated genomic sequence. It is expected that an increasing percentage of regions predicted to have function by process 200 will be assayed physically, and that display 80 will accordingly, for any given genomic sequence, have an increasing number of rectangles 84 and 85, representing an increased density of sequence annotation.

Where the function desired to be identified is protein coding, rectangle 84 identifies the sequence of the probe used to measure expression. In embodiments of the present invention where expression is measured using genome-derived single exon microarrays, rectangle 84 identifies the sequence included within the probe immobilized on the support surface of the microarray. As noted *supra*, such probe will often include a small amount of additional, synthetic, material incorporated during amplification and designed to permit reamplification of the

probe, which sequence is typically not shown in display 80.

Rectangle 87 is used to present the results of bioinformatic assay of the genomic sequence. For example, where the function desired to be identified is protein coding, process 400 can include bioinformatic query of expression databases with the sequences predicted in process 200 to encode exons. And as earlier discussed, because bioinformatic assay presents fewer constraints than does physical assay, often the entire output of process 200 can be used for such assay, without further subsetting thereof by process 300. Therefore, rectangle 87 typically need not have separate indicators therein of regions submitted for bioinformatic assay; that is, rectangle 87 typically need not have regions therein analogous to rectangles 84 within rectangle 89.

Rectangle 87 as shown in FIG. 3 includes smaller rectangles 880 and 88. Rectangles 880 indicate regions that returned a positive result in the bioinformatic assay, with rectangles 88 representing regions that did not return such positive results. Where the function desired to be predicted and displayed is protein coding, rectangles 880 indicate regions of the predicted exons that identify sequence with significant similarity in expression databases, such as EST, SNP, SAGE databases, with rectangles 88 indicating genes novel over those identified in existing expression data bases.

Rectangles 880 can further indicate, through color, shading, texture, or the like, additional information obtained from bioinformatic assay.

For example, where the function assayed and displayed is protein coding, the degree of shading of rectangles 880 can be used to represent the degree of sequence similarity found upon query of expression databases. The number of levels of discrimination can be as few as two (identity, and similarity, where similarity

has a user-selectable lower threshold). Alternatively, as many different levels of discrimination can be indicated as can visually be discriminated.

Where display 80 is used as a graphical user interface, rectangles 880 can additionally provide links directly to the sequences identified by the query of expression databases, and/or statistical summaries thereof. As with each of the precedingly-discussed uses of display 80 as a graphical user interface, it should be understood that the information accessed via display 80 need not be resident on the computer presenting such display, which often will be serving as a client, with the linked information resident on one or more remotely located servers.

Rectangle 85 displays the results of physical assay of the sequence delimited by its left and right borders.

Rectangle 85 can consist of a single rectangle, thus indicating a single assay, or alternatively, and increasingly typically, will consist of a series of rectangles (85a, 85b, 85c) indicating separate physical assays of the same sequence.

Where the function assayed is gene expression, and where gene expression is assayed as herein described using simultaneous two-color fluorescent detection of hybridization to genome-derived single exon microarrays, individual rectangles 85 can be colored to indicate the degree of expression relative to control. Conveniently, shades of green can be used to depict expression in the sample over control values, and shades of red used to depict expression less than control, corresponding to the spectra of the Cy3 and Cy5 dyes conventionally used for respective labeling thereof. Additional functional information can be provided in the form of circles 86 (86a, 86b, 86c), where the diameter of the circle can be used to

indicate expression intensity. As discussed *infra*, such relative expression (expression ratios) and absolute expression (signal intensity) can be expressed using normalized values.

Where display 80 is used as a graphical user interface, rectangle 85 can be used as a link to further information about the assay. For example, where the assay is one for gene expression, each rectangle 85 can be used to link to information about the source of the hybridized mRNA, the identity of the control, raw or processed data from the microarray scan, or the like.

FIG. 4 is rendition of display 80 representing gene prediction and gene expression for a hypothetical BAC, showing conventions used in the Examples presented *infra*. BAC sequence ("Chip seq.") 89 is presented, with the physically assayed region thereof (corresponding to rectangle 84 in FIG. 3) shown in white. Algorithmic gene predictions are shown in field 81, with predictions by GRAIL shown, predictions by GENEFINDER, and predictions by DICTION shown. Within rectangle 87, regions of sequence that, when used to query expression databases, return identical or similar sequences ("EST hit") are shown as white rectangles (corresponding to rectangles 880 in FIG. 3), gray indicates low homology, and black indicates unknowns (where black and gray would correspond to rectangles 88 in FIG. 3).

Although FIGS. 3 and 4 show a single stretch of sequence, uninterrupted from left to right, longer sequences are usefully represented by vertical stacking of such individual Mondrians, as shown in FIGS. 9 and 10.

Single Exon Probes Useful For Measuring Gene Expression

The methods and apparatus of the present invention rapidly produce functional information from

genomic sequence. Where the function to be identified is protein coding, the methods and apparatus of the present invention rapidly identify and confirm the expression of portions of genomic sequence that function to encode protein. As a direct result, the methods and apparatus of the present invention rapidly yield large numbers of single-exon nucleic acid probes, the majority from previously unknown genes, each of which is useful for measuring and/or surveying expression of a specific gene in one or more tissues or cell types.

It is, therefore, another aspect of the present invention to provide genome-derived single exon nucleic acid probes useful for gene expression analysis, and particularly for gene expression analysis by microarray.

Using the methods and genome-derived single-exon microarrays of the present invention, we have for example readily identified a large number of unique ORFs from human genomic sequence. Using single exon probes that encompass these ORFs, we have demonstrated, through microarray hybridization analysis, the expression of 13,109 of these ORFs in adult liver.

As would immediately be appreciated by one of skill in the art, each single exon probe having demonstrable expression in adult liver is currently available for use in measuring the level of its ORF's expression in adult liver.

Diseases of the liver are a significant cause of human morbidity and mortality. Increasingly, genetic factors are being found that contribute to predisposition, onset, and/or aggressiveness of most, if not all, of these diseases; although causative mutations in single genes have been identified for some, these disorders are believed for the most part to have polygenic etiologies.

For example, cirrhosis is a major public health problem. In the industrialized world, it is among the top

ten causes of death; among patients aged 45 to 65, it is the third leading cause of death. The high prevalence is largely the result of alcohol abuse, but other major contributors include chronic hepatitis, biliary disease and iron overload. Approximately 10-15% are cryptogenic.

Cirrhosis is a broad description encompassing the common end stage of many forms of liver injury. Many patients with cirrhosis will remain asymptomatic for years, while others show generalized weakness, anorexia, malaise, and weight loss or, occasionally, more severe symptoms.

The progression from fibrosis, an early consequence of liver disease, to cirrhosis, and the specific histologic morphology that characterizes cirrhosis depend on the extent of injury, the presence of continuing damage, and the response of the liver to damage. The liver may be injured acutely and severely (e.g. necrosis with hepatitis), moderately over months or years (e.g. biliary tract obstruction and chronic active hepatitis), or modestly but continuously (e.g. alcohol abuse).

During the repair process, new vessels connecting the hepatic artery and portal vein to the hepatic venules form within the fibrous sheath that surrounds the surviving nodules of liver cells. These vessels restore the intrahepatic circulatory pathway, but provide relatively low-volume, high-pressure drainage that is less efficient than normal and results in increased portal vein pressure (portal hypertension). Thus, cirrhosis is not static and its features depend on the disease activity and stage.

As cirrhosis is the end stage of many forms of liver disease, many genes have been identified that can contribute to the development of cirrhosis. These include, e.g., the genes responsible for Wilson disease (Online Mendelian Inheritance of Man ("OMIM") 277900), type IV glycogen storage disease (OMIM 232500), galactosemia (OMIM

230400), and a deficiency of alpha-1-antitrypsin (OMIM 107400). There is substantial evidence, however, for as yet uncharacterized loci which cause cirrhosis.

For example, Iber and Maddrey, Prog. Liver Dis. 2: 290-302 (1965), reviewed 13 previously reported families and 8 new to this study, each with 2 or more affected members. They pointed out that, with a single exception, the multiple cases were in the same generation. Within a given family, the age of onset, clinical course, and biopsy findings were very similar, but there were wide differences between families.

Kalra et al., Hum. Hered. 32:170-175 (1982) studied the families of 220 cases of Indian childhood cirrhosis and 70 families of age-matched controls. The hypotheses of autosomal recessive, partial sex-linkage, and doubly recessive inheritance were found untenable and the authors concluded that multifactorial inheritance was most plausible. Lefkowitch et al., New Eng. J. Med. 307:271-277 (1982) described 4 white American sibs who died between ages 4.5 and 6 years of cirrhosis that closely resembled that of the childhood cirrhosis of Asiatic Indians.

Another example of uncharacterized loci which cause cirrhosis are those related to the risk of alcoholism.

Cloninger, Science 236:410-416 (1987), defined two separate types of alcoholism. According to these definitions, type 1 alcohol abuse has its usual onset after the age of 25 years and is characterized by severe psychological dependence and guilt. Type 1 occurs in both men and women and requires both genetic and environmental factors to become manifest. By contrast, type 2 alcohol abuse has its onset before the age of 25; persons with this type of alcoholism are characterized by their inability to abstain from alcohol and by frequent aggressive and antisocial behavior. Type 2 alcoholism is rarely found in

women and is much more heritable.

Despite considerable effort to identify genes related to the risk of alcoholism, relatively few genes have been identified. Some of this work has suggested a relationship between the metabolism of dopamine and alcoholism. Blum et al., J.A.M.A. 263:2055-2060 (1990) and Bolos et al., J.A.M.A. 264:3156-3160 (1990) investigated the relationship of the dopamine D2 receptor (DRD2; OMIM 126450) to alcoholism, but the sample size was small and their results were inconclusive. However, Tiihonen et al., Molec. Psychiat. 4, 286-289 (1999), found a markedly higher frequency in a population of type 1 alcoholics of the low activity allele of the enzyme catechol-O-methyltransferase (COMT, OMIM 116790), which has a crucial role in the metabolism of dopamine, suggesting a role for dopamine metabolism in increased risk of alcoholism. For a brief review of recent progress toward the identification of genes related to risk for alcoholism see Buck, Genome 9:927-928 (1998).

As another example, multiple genes have been shown to predispose to hyperlipoproteinemia or hyperlipidemia. Much attention has been focused on these disorders because there is a strong association of hyperlipidemia, especially hypercholesterolemia, with development of coronary artery disease. Coronary artery disease accounts for at least 25% of all deaths in the United States. Coronary artery disease results when the arteries supplying the heart muscle become occluded by plaques composed of lipids like cholesterol, blood clotting components and blood cells.

The major plasma lipids circulate bound to proteins as macromolecular complexes called lipoproteins. Although closely interrelated, the major lipoprotein classes - chylomicron, very-low-density lipoprotein (VLDL), low-density lipoprotein (LDL), and high-density lipoprotein

(HDL) - are usually classified in terms of physicochemical properties (e.g., density after centrifugation). Chylomicrons, the largest lipoproteins, carry exogenous triglyceride from the intestine via the thoracic duct to the venous system and into peripheral sites. VLDL carries endogenous triglyceride primarily from the liver to the same peripheral sites for storage or use. Lipases quickly degrade the triglyceride in VLDL to produce intermediate density lipoproteins (IDL) and within 2 to 6 h, IDL is degraded further to generate LDL, which has a plasma half-life of 2 to 3 days. While the overall fate of LDL is unclear, the liver is responsible for removing approximately 70% and active receptor sites have been found on the surfaces of hepatocytes.

Several monogenic conditions that lead to elevated levels of one or more serum lipoproteins have been defined and the responsible gene identified, including, e.g., hyperlipoproteinemia type I (OMIM 238600), familial hypercholesterolemia (OMIM 143890), and familial defective apolipoprotein B (OMIM 107730). However, in many cases the etiology is unknown and there is strong evidence for additional uncharacterized loci.

For example, Zuliani et al., Arterioscler. Thromb. Vasc. Biol. 19:802-809 (1999) identified a Sardinian family with a recessive form of hypercholesterolemia with the clinical features of familial hypercholesterolemia (OMIM 603813), and found that previously identified genes were not responsible for this disorder. They proposed that in this new lipid disorder, a recessive defect causes a selective impairment of the LDL receptor function in the liver. Ciccarese et al., Am. J. Hum. Genet. 66:453-460 (2000) recently mapped this novel disease locus.

Another example is designated familial combined hyperlipidemia (OMIM 144250) which affects approximately 1-

2% of the population in the Western world. This disorder can have its basis in mutation in several novel genes, two of which have been mapped to chromosome 1 (Pajukanta et al., Nature Genet. 18:369-373 (1998)) and chromosome 11 (Aouizerat et al., Am. J. Hum. Genet. 65, 397-412 (1999)). The high frequency of this disorder suggests that most, if not all, hyperlipidemias are of multifactorial genetic etiology.

As yet a further example, primary schlerosing cholangitis (PSC) is a disorder characterized by a patchy obliterative inflammatory fibrosis of the large bile ducts. Chronic inflammation leads to extensive bile duct strictures, cholestasis, and gradual progression to biliary cirrhosis. PSC occurs most often in young men and is commonly associated with inflammatory bowel disease, especially ulcerative colitis. The onset is usually insidious, with gradual, progressive fatigue, pruritus, and jaundice. There is no specific therapy for sclerosing cholangitis, and liver transplantation is the only apparent cure.

The etiology of PSC is not known, but both genetic and immunologic abnormalities have been implicated. However, the frequency of HLA-B8 and HLA-DT2, which are associated with a number of autoimmune diseases, is higher in PSC than normal individuals. Prochazka et al., New Eng. J. Med. 322:1842-1844 (1990) found that 100% of 29 patients with primary sclerosing cholangitis carried the HLA-DRw52a antigen, which is normally present in 35% of the population.

As a still further example, sarcoidosis is a disease of unknown cause characterized by non-caseating granulomas in one or more organ systems. These granulomas may resolve completely or proceed to fibrosis. The disorder is systemic, but the liver is affected in approximately 75% of cases. Sarcoidosis occurs mainly in persons aged 20 to

40 yr and is most common in Northern Europeans and American blacks. The lifetime risk of developing sarcoidosis is particularly high among Swedish men (1.15%), Swedish women (1.6%), and African Americans (2.4%).

The much greater frequency in African Americans relative to the United States population overall suggests a genetic contribution to etiology. Early research studying familial aggregation indicated that the disease may have a nongenetic basis because the family pattern did not conform to a simple Mendelian mode of inheritance (Allison, Sth. Med. J. 57: 27-32 (1964)). However, Headings et al., Ann. N.Y. Acad. Sci. 278:377-385 (1976) favored multifactorial genetic inheritance of susceptibility. Nowack et al., Arch. Intern. Med. 147:481-483 (1987), found an unusually high frequency of HLA-DR5 in a study of 440 patients with sarcoidosis in Marburg, Germany. They also concluded that the role of an environmental or infectious agent triggering sarcoidosis cannot be envisaged without considering genetically linked cofactors.

Other significant diseases of liver are also believed to have a genetic, typically polygenic, etiologic component. These diseases include, e.g., primary biliary cirrhosis, Zellweger syndrome, cholestasis-lymphedema syndrome, Alstrom syndrome, primary pulmonary hypertension, Berardinelli-Seip congenital lipodystrophy, iron overload in Africa, neonatal cholestatic hepatitis, autosomal recessive KID syndrome, familial hypotransferrinemia, type I congenital dyserythropoietic anemia, porphyria variegata, Finnish lactic acidosis with hepatic hemosiderosis, Rotor syndrome, essential hypertension, ARC syndrome, type II conjugated hyperbilirubinemia, Lambert syndrome, ichthyosis congenita with biliary atresia, Kabuki make-up syndrome, Meckel syndrome, cerebral aneurysm-cirrhosis syndrome, glycogen storage diseases, polycystic kidney and hepatic disease,

isolated Caroli disease, trisomy 18-like syndrome, Osler-Rendu-Weber syndrome 3, fatal intrahepatic cholestasis, Coach syndrome, type C Niemann-Pick disease, and hepatocellular cancer.

Altered responses to a variety of infectious agents that target the liver, especially acute viral hepatitis, have also been shown or are suspected to have genetic bases or contributions. In addition to differential susceptibility to primary infectious agents, these altered responses include predisposition to complicating conditions following contact with particular infectious agents. These include, e.g., development of hepatocellular carcinoma 2 correlated with Hepatitis B infection, and severe hepatic fibrosis following Schistosoma mansoni infection.

The central role of the liver in drug metabolism results in exposure of this organ to a large variety of potentially toxic chemical agents and metabolites. These include naturally occurring plant alkaloids and mycotoxins, industrial chemicals, and, additionally, pharmacologic agents used in treating disease. The range of manifestations of toxin- and drug-induced liver disease are virtually as broad as the range of acute and chronic disorders and have also been shown or suspected to have genetic bases or contributions.

Such interactions between drugs and genotype have been shown in the response, e.g., to the anticonvulsant phenytoin, which can cause severe hepatitis-like disease in individuals who are impaired in the ability to detoxify a metabolite of phenytoin in the liver, and in the response to the drug sodium valproate, which can produce severe hepatotoxicity in certain individuals. The abnormal responses to both of these drugs are believed to be influenced by underlying genetic factors.

The human genome-derived single exon nucleic acid

probes and microarrays of the present invention are useful for predicting, diagnosing, grading, staging, monitoring and prognosing diseases of human liver, particularly those diseases with polygenic etiology. With each of the single exon probes described herein shown to be expressed at detectable levels in human liver, and with about 2/3 of the probes identifying novel genes, the single exon microarrays of the present invention provide exceptionally high informational content for such studies.

For example, diagnosis (including differential diagnosis among clinically indistinguishable disorders, such as cirrhosis), staging, and/or grading of a disease can be based upon the quantitative relatedness of a patient gene expression profile to one or more reference expression profiles known to be characteristic of a given liver disease, or to specific grades or stages thereof.

In one embodiment, the patient gene expression profile is generated by hybridizing nucleic acids obtained directly or indirectly from transcripts expressed in the patient's liver to the genome-derived single exon microarray of the present invention. Reference profiles are obtained similarly, using nucleic acids obtained directly or indirectly from transcripts expressed by liver of individuals with known liver disease. Methods for quantitatively relating gene expression profiles, without regard to the function of the protein encoded by the gene, are disclosed in WO 99/58720, incorporated herein by reference in its entirety.

In another approach, the genome-derived single exon probes and microarrays of the present invention can be used to interrogate genomic DNA, rather than pools of expressed message; this latter approach permits predisposition to and/or prognosis of liver disease to be assessed through the massively parallel determination of altered copy number, deletion, or mutation in the patient's

genome of exons known to be expressed in human liver. The algorithms set forth in WO 99/58720 can be applied to such genomic profiles without regard to the function of the protein encoded by the interrogated gene.

The utility is specific to the probe; at sufficiently high hybridization stringency, which stringencies are well known in the art — see Ausubel et al. and Maniatis *et al.* — each probe reports the level of expression of message specifically containing that ORF.

It should be appreciated, however, that the probes of the present invention, for which expression in the adult liver has been demonstrated are useful for both measurement in the adult liver and for survey of expression in other tissues.

Significant among such advantages is the presence of probes for novel genes.

As mentioned above and further detailed in Examples 1 and 2, the methods described enable ORFs which are not present in existing expression databases to be identified. And the fewer the number of tissues in which the ORF can be shown to be expressed, the more likely the ORF will prove to be part of a novel gene: as further discussed in Example 2, ORFs whose expression was measurable in only a single of the tested tissues were represented in existing expression databases at a rate of only 11%, whereas 36% of ORFs whose expression was measurable in 9 tissues were present in existing expression databases, and fully 45% of those ORFs expressed in all ten tested tissues were present in existing expressed sequence databases.

Either as tools for measuring gene expression or tools for surveying gene expression, the genome-derived single exon probes of the present invention have significant advantages over the cDNA or EST-based probes that are currently available for achieving these utilities.

The genome-derived single exon probes of the present invention are useful in constructing genome-derived single exon microarrays; the genome-derived single exon microarrays, in turn, are useful devices for measuring and for surveying gene expression in the human.

Gene expression analysis using microarrays — conventionally using microarrays having probes derived from expressed message — is well-established as useful in the biological research arts (see Lockhart et al. Nature 405, 827-836).

Microarrays have been used to determine gene expression profiles in cells in response to drug treatment (see, for example, Kaminski *et al.*, "Global Analysis of Gene Expression in Pulmonary Fibrosis Reveals Distinct Programs Regulating Lung Inflammation and Fibrosis," *Proc. Natl. Acad. Sci. USA* 97(4):1778-83 (2000); Bartosiewicz *et al.*, "Development of a Toxicological Gene Array and Quantitative Assessment of This Technology," *Arch. Biochem. Biophys.* 376(1):66-73 (2000)), viral infection (see for example, Geiss *et al.*, "Large-scale Monitoring of Host Cell Gene Expression During HIV-1 Infection Using cDNA Microarrays," *Virology* 266(1):8-16 (2000)) and during cell processes such as differentiation, senescence and apoptosis (see, for example, Shelton *et al.*,"Microarray Analysis of Replicative Senescence," *Curr. Biol.* 9(17):939-45 (1999); Voehringer *et al.*, "Gene Microarray Identification of Redox and Mitochondrial Elements That Control Resistance or Sensitivity to Apoptosis," Proc. Natl. Acad. Sci. USA 97(6):2680-5 (2000)).

Microarrays have also been used to determine abnormal gene expression in diseased tissues (see, for example, Alon *et al.*, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Natl. Acad. Sci. USA* 96(12):6745-50 (1999); Perou *et al.*,

"Distinctive Gene Expression Patterns in Human Mammary Epithelial Cells and Breast Cancers, *Proc. Natl. Acad. Sci. USA* 96(16):9212-7 (1999); Wang *et al.*, "Identification of Genes Differentially Over-expressed in Lung Squamous Cell Carcinoma Using Combination of cDNA Subtraction and Microarray Analysis," *Oncogene* 19(12):1519-28 (2000); Whitney *et al.*, "Analysis of Gene Expression in Multiple Sclerosis Lesions Using cDNA Microarrays," *Ann. Neurol.* 46(3):425-8 (1999)), in drug discovery screens (see, for example, Scherf *et al.*, "A Gene Expression Database for the Molecular Pharmacology of Cancer," *Nat. Genet.* 24(3):236-44 (2000)) and in diagnosis to determine appropriate treatment strategies (see, for example, Sgroi *et al.*, "*In vivo* Gene Expression Profile Analysis of Human Breast Cancer Progression," *Cancer Res.* 59(22):5656-61 (1999)).

In microarray-based gene expression screens of pharmacological drug candidates upon cells, each probe provides specific useful data. In particular, it should be appreciated that even those probes that show no change in expression are as informative as those that do change, serving, in essence, as negative controls.

For example, where gene expression analysis is used to assess toxicity of chemical agents on cells, the failure of the agent to change a gene's expression level is evidence that the drug likely does not affect the pathway of which the gene's expressed protein is a part. Analogously, where gene expression analysis is used to assess side effects of pharmacological agents – whether in lead compound discovery or in subsequent screening of lead compound derivatives – the inability of the agent to alter a gene's expression level is evidence that the drug does not affect the pathway of which the gene's expressed protein is a part.

WO 99/58720 provides methods for quantifying the relatedness of a first and second gene expression profile

and for ordering the relatedness of a plurality of gene expression profiles. The methods so described permit useful information to be extracted from a greater percentage of the individual gene expression measurements from a microarray than methods previously used in the art.

Other uses of microarrays are described in Gerhold *et al., Trends Biochem. Sci.* 24(5):168-173 (1999) and Zweiger, *Trends Biotechnol.* 17(11):429-436 (1999); Schena et al.

The invention particularly provides genome-derived single-exon probes known to be expressed in adult liver. The individual single exon probes can be provided in the form of substantially isolated and purified nucleic acid, typically, but not necessarily, in a quantity sufficient to perform a hybridization reaction.

Such nucleic acid can be in any form directly hybridizable to the message that contains the probe's ORF, such as double stranded DNA, single-stranded DNA complementary to the message, single-stranded RNA complementary to the message, or chimeric DNA/RNA molecules so hybridizable. The nucleic acid can alternatively or additionally include either nonnative nucleotides, alternative internucleotide linkages, or both, so long as complementary binding can be obtained. For example, probes can include phosphorothioates, methylphosphonates, morpholino analogs, and peptide nucleic acids (PNA), as are described, for example, in U.S. Patent Nos. 5,142,047; 5,235,033; 5,166,315; 5,217,866; 5,184,444; 5,861,250.

Usefully, however, such probes are provided in a form and quantity suitable for amplification, where the amplified product is thereafter to be used in the hybridization reactions that probe gene expression. Typically, such probes are provided in a form and quantity suitable for amplification by PCR or by other well known amplification technique. One such technique additional to

PCR is rolling circle amplification, as is described, *inter alia*, in U.S. Patent Nos. 5,854,033 and 5,714,320 and international patent publications WO 97/19193 and WO 00/15779. As is well understood, where the probes are to be provided in a form suitable for amplification, the range of nucleic acid analogues and/or internucleotide linkages will be constrained by the requirements and nature of the amplification enzyme.

Where the probe is to be provided in form suitable for amplification, the quantity need not be sufficient for direct hybridization for gene expression analysis, and need be sufficient only to function as an amplification template, typically at least about 1, 10 or 100 pg or more.

Each discrete amplifiable probe can also be packaged with amplification primers, either in a single composition that comprises probe template and primers, or in a kit that comprises such primers separately packaged therefrom. As earlier mentioned, the ORF-specific 5' primers used for genomic amplification can have a first common sequence added thereto, and the ORF-specific 3' primers used for genomic amplification can have a second, different, common sequence added thereto, thus permitting, in this embodiment, the use of a single set of 5' and 3' primers to amplify any one of the probes. The probe composition and/or kit can also include buffers, enzyme, *etc.*, required to effect amplification.

As mentioned earlier, when intended for use on a genome-derived single exon microarray of the present invention, the genome-derived single exon probes of the present invention will typically average at least about 100, 200, 300, 400 or 500 bp in length, including (and typically, but not necessarily centered about) the ORF. Furthermore, when intended for use on a genome-derived single exon microarray of the present invention, the

genome-derived single exon probes of the present invention will typically not contain a detectable label.

When intended for use in solution phase hybridization, however – that is, for use in a hybridization reaction in which the probe is not first bound to a support substrate (although the target may indeed be so bound) – length constraints that are imposed in microarray-based hybridization approaches will be relaxed, and such probes will typically be labeled.

In such case, the only functional constraint that dictates the minimum size of such probe is that each such probe must be capable of specifically identifying in a hybridization reaction the exon from which it is drawn. In theory, a probe of as little as 17 nucleotides is capable of uniquely identifying its cognate sequence in the human genome. For hybridization to expressed message – a subset of target sequence that is much reduced in complexity as compared to genomic sequence – even fewer nucleotides are required for specificity.

Therefore, the probes of the present invention can include as few as 20, 25 or 50 bp or ORF, or more. In particular embodiments, the ORF sequences are given in SEQ ID NOS. 13,110 - 25,995, respectively, for probe SEQ ID NOS. 1 - 13,109. The minimum amount of ORF required to be included in the probe of the present invention in order to provide specific signal in either solution phase or microarray-based hybridizations can readily be determined for each of ORF SEQ ID NOS. 13,110 - 25,995 individually by routine experimentation using standard high stringency conditions.

Such high stringency conditions are described, *inter alia*, in Ausubel et al. and Maniatis *et al.* For microarray-based hybridization, standard high stringency conditions can usefully be 50% formamide, 5X SSC, 0.2 µg/µl poly(dA), 0.2 µg/µl human $c_ot1$ DNA, and 0.5 % SDS, in a

humid oven at 42°C overnight, followed by successive washes of the microarray in 1X SSC, 0.2% SDS at 55°C for 5 minutes, and then 0.1X SSC, 0.2% SDS, at 55°C for 20 minutes. For solution phase hybridization, standard high stringency conditions can usefully be aqueous hybridization at 65°C in 6X SSC. Lower stringency conditions, suitable for cross-hybridization to mRNA encoding structurally- and functionally-related proteins, can usefully be the same as the high stringency conditions but with reduction in temperature for hybridization and washing to room temperature (approximately 25°C).

When intended for use in solution phase hybridization, the maximum size of the single exon probes of the present invention is dictated by the proximity of other expressed exons in genomic DNA: although each single exon probe can include intergenic and/or intronic material contiguous to the ORF in the human genome, each probe of the present invention will include portions of only one expressed exon.

Thus, each single exon probe will include no more than about 25 kb of contiguous genomic sequence, more typically no more than about 20 kb of contiguous genomic sequence, more usually no more than about 15 kb, even more usually no more than about 10 kb. Usually, probes that are maximally about 5 kb will be used, more typically no more than about 3 kb.

It will be appreciated that the Sequence Listing appended hereto presents, by convention, only that strand of the probe and ORF sequence that can be directly translated reading from 5' to 3' end. As would be well understood by one of skill in the art, single stranded probes must be complementary in sequence to the ORF as present in an mRNA; it is well within the skill in the art to determine such complementary sequence. It will further be understood that double stranded probes can be used in

both solution-phase hybridization and microarray-based hybridization if suitably denatured.

Thus, it is an aspect of the present invention to provide single-stranded nucleic acid probes that have sequence complementary to those described herein above and below, and double-stranded probes one strand of which has sequence complementary to the probes described herein.

The probes can, but need not, contain intergenic and/or intronic material that flanks the ORF, on one or both sides, in the same linear relationship to the ORF that the intergenic and/or intronic material bears to the ORF in genomic DNA. The probes do not, however, contain nucleic acid derived from more than one expressed ORF.

And when intended for use in solution hybridization, the probes of the present invention can usefully have detectable labels. Nucleic acid labels are well known in the art, and include, *inter alia*, radioactive labels, such as $^{3}H$, $^{32}P$, $^{33}P$, $^{35}S$, $^{125}I$, $^{131}I$; fluorescent labels, such as Cy3, Cy5, Cy5.5, Cy7, SYBR®

Green and other labels described in Haugland, *Handbook of Fluorescent Probes and Research Chemicals*, 7th ed., Molecular Probes Inc., Eugene, OR (2000), or fluorescence resonance energy transfer tandem conjugates thereof; labels suitable for chemiluminescent and/or enhanced chemiluminescent detection; labels suitable for ESR and NMR detection; and labels that include one member of a specific binding pair, such as biotin, digoxigenin, or the like.

The probes, either in quantity sufficient for hybridization or sufficient for amplification, can be provided in individual vials or containers.

Alternatively, such probes can usefully be packaged as a plurality of such individual genome-derived single exon probes.

When provided as a collection of plural

individual probes, the probes are typically made available in amplifiable form in a spatially-addressable ordered set, typically one per well of a microtiter dish. Although a 96 well microtiter plate can be used, greater efficiency is obtained using higher density arrays.

If, as earlier mentioned, the ORF-specific 5' primers used for genomic amplification had a first common sequence added thereto, and the ORF-specific 3' primers used for genomic amplification had a second, different, common sequence added thereto, a single set of 5' and 3' primers can be used to amplify all of the probes from the amplifiable ordered set.

Such collections of genome-derived single exon probes can usefully include a plurality of probes chosen for the common attribute of expression in the human adult liver.

In such defined subsets, typically at least 50, 60, 75, 80, 85, 90 or 95% or more of the probes will be chosen by their expression in the defined tissue or cell type.

The single exon probes of the present invention, as well as fragments of the single exon probes comprising selectively hybridizable portions of the probe ORF, can be used to obtain the full length cDNA that includes the ORF by (i) screening of cDNA libraries; (ii) rapid amplification of cDNA ends ("RACE"); or (iii) other conventional means, as are described, *inter alia*, in Ausubel et al. and Maniatis *et al.*

It is another aspect of the present invention to provide genome-derived single exon nucleic acid microarrays useful for gene expression analysis, where the term "microarray" has the meaning given in the definitional section of this description, *supra*.

The invention particularly provides genome-derived single-exon nucleic acid microarrays comprising a

plurality of probes known to be expressed in human adult liver. In preferred embodiments, the present invention provides human genome-derived single exon microarrays comprising a plurality of probes drawn from the group consisting of SEQ ID NOS.: 1 - 13,109.

When used for gene expression analysis, the genome-derived single exon microarrays provide greater physical informational density than do the genome-derived single exon microarrays that have lower percentages of probes known to be expressed commonly in the tested tissue. At a fixed probe density, for example, a given microarray surface area of the defined subset genome-derived single exon microarray can yield a greater number of expression measurements. Alternatively, at a given probe density, the same number of expression measurements can be obtained from a smaller substrate surface area. Alternatively, at a fixed probe density and fixed surface area, probes can be provided redundantly, providing greater reliability in signal measurement for any given probe. Furthermore, with a higher percentage of probes known to be expressed in the assayed tissue, the dynamic range of the detection means can be adjusted to reveal finer levels discrimination among the levels of expression.

Although particularly described with respect to their utility as probes of gene expression, particularly as probes to be included on a genome-derived single exon microarray, each of the nucleic acids having SEQ ID NOS.: 1 - 13,109contains an open-reading frame, set forth respectively in SEQ ID NOS.: 13,110 - 25,995, that encodes a protein domain. Thus, each of SEQ ID NOS. 1 - 13,109can be used, or that portion thereof in SEQ ID NOS. 13,110 - 25,995 used, to express a protein domain by standard *in vitro* recombinant techniques. See Ausubel et al. and Maniatis *et al.*

Additionally, kits are available commercially

that readily permit such nucleic acids to be expressed as protein in bacterial cells, insect cells, or mammalian cells, as desired (*e.g.*, HAT™ Protein Expression & Purification System, ClonTech Laboratories, Palo Alto, CA; Adeno-X™ Expression System, ClonTech Laboratories, Palo Alto, CA; Protein Fusion & Purification (pMAL™) System, New England Biolabs, Beverley, MA)

Furthermore, shorter peptides can be chemically synthesized using commercial peptide synthesizing equipment and well known techniques. Procedures are described, *inter alia*, in Chan *et al.* (eds.), Fmoc Solid Phase Peptide Synthesis: A Practical Approach (Practical Approach Series, (Paper)), Oxford Univ. Press (March 2000) (ISBN: 0199637245); Jones, Amino Acid and Peptide Synthesis (Oxford Chemistry Primers, No 7) , Oxford Univ. Press (August 1992) (ISBN: 0198556683); and Bodanszky, Principles of Peptide Synthesis (Springer Laboratory), Springer Verlag (December 1993) (ISBN: 0387564314).

It is, therefore, another aspect of the invention to provide peptides comprising an amino acid sequence translated from SEQ ID NOS.: 13,110 - 25,995. Such amino acid sequences are set out in SEQ ID NOS: 25,996 - 38,578. Any such recombinantly-expressed or synthesized peptide of at least 8, and preferably at least about 15, amino acids, can be conjugated to a carrier protein and used to generate antibody that recognizes the peptide. Thus, it is a further aspect of the invention to provide peptides that have at least 8, preferably at least 15, consecutive amino acids.

The following examples are offered by way of illustration and not by way of limitation.

EXAMPLE 1

Preparation of Single Exon Microarrays from ORFs Predicted

in Human Genomic Sequence

Bioinformatics Results

All human BAC sequences in fewer than 10 pieces that had been accessioned in a five month period immediately preceding this study were downloaded from GenBank. This corresponds to ~2200 clones, totaling ~350 MB of sequence, or approximately 10% of the human genome.

After masking repetitive elements using the program CROSS_MATCH, the sequence was analyzed for open reading frames using three separate gene finding programs. The three programs predict genes using independent algorithmic methods developed on independent training sets: GRAIL uses a neural network, GENEFINDER uses a hidden Markoff model, and DICTION, a program proprietary to Genetics Institute, operates according to a different heuristic. The results of all three programs were used to create a prediction matrix across the segment of genomic DNA.

The three gene finding programs yielded a range of results. GRAIL identified the greatest percentage of genomic sequence as putative coding region, 2% of the data analyzed. GENEFINDER was second, calling 1%, and DICTION yielded the least putative coding region, with 0.8% of genomic sequence called as coding region.

The consensus data were as follows. GRAIL and GENEFINDER agreed on 0.7% of genomic sequence, GRAIL and DICTION agreed on 0.5% of genomic sequence, and the three programs together agreed on 0.25% of the data analyzed. That is, 0.25% of the genomic sequence was identified by all three of the programs as containing putative coding region.

ORFs predicted by any two of the three programs ("consensus ORFs") were assorted into "gene bins" using two criteria: (1) any 7 consecutive exons within a 25 kb window

were placed together in a bin as likely contributing to a single gene, and (2) all ORFs within a 25 kb window were placed together in a bin as likely contributing to a single gene if fewer than 7 exons were found within the 25 kb window.

PCR

The largest ORF from each gene bin that did not span repetitive sequence was then chosen for amplification, as were all consensus ORFs longer than 500 bp. This method approximated one exon per gene; however, a number of genes were found to be represented by multiple elements.

Previously, we had determined that DNA fragments fewer than 250 bp in length do not bind well to the amino-modified glass surface of the slides used as support substrate for construction of microarrays; therefore, amplicons were designed in the present experiments to approximate 500 bp in length.

Accordingly, after selecting the largest ORF per gene bin, a 500 bp fragment of sequence centered on the ORF was passed to the primer picking software, PRIMER3 (available online for use at http://www-genome.wi.mit.edu/cgi-bin/primer/ ). A first additional sequence was commonly added to each ORF-unique 5' primer, and a second, different, additional sequence was commonly added to each ORF-unique 3' primer, to permit subsequent reamplification of the amplicon using a single set of "universal" 5' and 3' primers, thus immortalizing the amplicon. The addition of universal priming sequences also facilitates sequence verification, and can be used to add a cloning site should some ORFs be found to warrant further study.

The ORFs were then PCR amplified from genomic DNA, verified on agarose gels, and sequenced using the universal primers to validate the identity of the amplicon

to be spotted in the microarray.

Primers were supplied by Operon Technologies (Alameda, CA). PCR amplification was performed by standard techniques using human genomic DNA (Clontech, Palo Alto, CA) as template. Each PCR product was verified by SYBR® green (Molecular Probes, Inc., Eugene, OR) staining of agarose gels, with subsequent imaging by Fluorimager (Molecular Dynamics, Inc., Sunnyvale, CA). PCR amplification was classified as successful if a single band appeared.

The success rate for amplifying ORFs of interest directly from genomic DNA using PCR was approximately 75%. FIG. 5 graphs the distribution of predicted ORF (exon) length and distribution of amplified PCR products, with ORF length shown in red and PCR product length shown in blue (which may appear black in the figure). Although the range of ORF sizes is readily seen to extend to beyond 900 bp, the mean predicted exon size was only 229 bp, with a median size of 150 bp (n=9498). With an average amplicon size of 475 ± 25 bp, approximately 50% of the average PCR amplification product contained predicted coding region, with the remaining 50% of the amplicon containing either intron, intergenic sequence, or both.

Using a strategy predicated on amplifying about 500 bp, it was found that long exons had a higher PCR failure rate. To address this, the bioinformatics process was adjusted to amplify 1000, 1500 or 2000 bp fragments from exons larger than 500 bp. This improved the rate of successful amplification of exons exceeding 500 bp, constituting about 9.2% of the exons predicted by the gene finding algorithms.

Approximately 75% of the probes disposed on the array (90% of those that successfully PCR amplified) were sequence-verified by sequencing in both the forward and reverse direction using MegaBACE sequencer (Molecular

Dynamics, Inc., Sunnyvale, CA), universal primers, and standard protocols.

Some genomic clones (BACs) yielded very poor PCR and sequencing results. The reasons for this are unclear, but may be related to the quality of early draft sequence or the inclusion of vector and host contamination in some submitted sequence data.

Although the intronic and intergenic material flanking coding regions could theoretically interfere with hybridization during microarray experiments, subsequent empirical results demonstrated that differential expression ratios were not significantly affected by the presence of noncoding sequence. The variation in exon size was similarly found not to affect differential expression ratios significantly; however, variation in exon size was observed to affect the absolute signal intensity (data not shown).

The 350 MB of genomic DNA was, by the above-described process, reduced to 9750 discrete probes, which were spotted in duplicate onto glass slides using commercially available instrumentation (MicroArray GenII Spotter and/or MicroArray GenIII Spotter, Molecular Dynamics, Inc., Sunnyvale, CA). Each slide additionally included either 16 or 32 *E. coli* genes, the average hybridization signal of which was used as a measure of background biological noise.

Each of the probe sequences was BLASTed against the human EST data set, the NR data set, and SwissProt GenBank (May 7, 1999 release 2.0.9).

One third of the probe sequences (as amplified) produced an exact match (BLAST Expect ("E") values less than 1 $e^{-100}$) to either an EST (20% of sequences) or a known mRNA (13% of sequences). A further 22% of the probe sequences showed some homology to a known EST or mRNA (BLAST E values from 1 $e^{-5}$ to 1 $e^{-99}$). The remaining 45% of

the probe sequences showed no significant sequence homology to any expressed, or potentially expressed, sequences present in public databases.

All of the probe sequences (as amplified) were then analyzed for protein similarities with the SwissProt database using BLASTX, Gish *et al.*, *Nature Genet.* 3:266 (1993). The predicted functional breakdowns of the 2/3 of probes identical or homologous to known sequences are presented in Table 1.

Table 1

| Function of Predicted ORFs As Deduced From Comparative Sequence Analysis | | | |
|---|---|---|---|
| Total | V6 chip | V7 chip | Function Predicted from Comparative Sequence Analysis |
| 211 | 96 | 115 | Receptor |
| 120 | 43 | 77 | Zinc Finger |
| 30 | 11 | 19 | Homeobox |
| 25 | 9 | 16 | Transcription Factor |
| 17 | 11 | 7 | Transcription |
| 118 | 57 | 61 | Structural |
| 95 | 39 | 56 | Kinase |
| 36 | 18 | 18 | Phosphatase |
| 83 | 31 | 52 | Ribosomal |
| 45 | 19 | 26 | Transport |
| 21 | 17 | 14 | Growth Factor |
| 17 | 12 | 5 | Cytochrome |
| 50 | 33 | 17 | Channel |

As can be seen, the two most common types of genes were transcription factors and receptors, making up 2.2% and 1.8% of the arrayed elements, respectively.

EXAMPLE 2

Gene Expression Measurements From Genome-Derived Single Exon Microarrays

The two genome-derived single exon microarrays prepared according to Example 1 were hybridized in a series of simultaneous two-color fluorescence experiments to (1) Cy3-labeled cDNA synthesized from message drawn individually from each of brain, heart, liver, fetal liver, placenta, lung, bone marrow, HeLa, BT 474, or HBL 100 cells, and (2) Cy5-labeled cDNA prepared from message pooled from all ten tissues and cell types, as a control in each of the measurements. Hybridization and scanning were carried out using standard protocols and Molecular Dynamics equipment.

Briefly, mRNA samples were bought from commercial sources (Clontech, Palo Alto, CA and Amersham Pharmacia Biotech (APB)). Cy3-dCTP and Cy5-dCTP (both from APB) were incorporated during separate reverse transcriptions of 1 µg of polyA$^+$ mRNA performed using 1 µg oligo(dT)12-18 primer and 2 µg random 9mer primers as follows. After heating to 70°C, the RNA:primer mixture was snap cooled on ice. After snap cooling on ice, added to the RNA to the stated final concentration was: 1X Superscript II buffer, 0.01 M DTT, 100µM dATP, 100 µM dGTP, 100 µM dTTP, 50 µM dCTP, 50 µM Cy3-dCTP or Cy5-dCTP 50 µM, and 200 U Superscript II enzyme. The reaction was incubated for 2 hours at 42°C. After 2 hours, the first strand cDNA was isolated by adding 1 U Ribonuclease H, and incubating for 30 minutes at 37°C. The reaction was then purified using a Qiagen PCR cleanup column, increasing the number of ethanol washes to 5. Probe was eluted using 10 mM Tris pH 8.5.

Using a spectrophotometer, probes were measured for dye incorporation. Volumes of both Cy3 and Cy5 cDNA corresponding to 50 pmoles of each dye were then dried in a Speedvac, resuspended in 30 µl hybridization solution containing 50% formamide, 5X SSC, 0.2 µg/µl poly(dA), 0.2 µg/µl human $c_ot1$ DNA, and 0.5 % SDS.

Hybridizations were carried out under a coverslip, with the array placed in a humid oven at 42°C overnight. Before scanning, slides were washed in 1X SSC, 0.2% SDS at 55°C for 5 minutes, followed by 0.1X SSC, 0.2% SDS, at 55°C for 20 minutes. Slides were briefly dipped in water and dried thoroughly under a gentle stream of nitrogen.

Slides were scanned using a Molecular Dynamics Gen3 scanner, as described. Schena (ed.), Microarray Biochip: Tools and Technology, Eaton Publishing Company/BioTechniques Books Division (2000) (ISBN: 1881299376).

Although the use of pooled cDNA as a reference permitted the survey of a large number of tissues, it attenuates the measurement of relative gene expression, since every highly expressed gene in the tissue/cell type-specific fluorescence channel will be present to a level of at least 10% in the control channel. Because of this fact, both signal and expression ratios (the latter hereinafter, "expression" or "relative expression") for each probe were normalized using the average ratio or average signal, respectively, as measured across the whole slide.

Data were accepted for further analysis only when signal was at least three times greater than biological noise, the latter defined by the average signal produced by the *E. coli* control genes.

The relative expression signal for these probes was then plotted as function of tissue or cell type, and is presented in FIG. 6.

FIG. 6 shows the distribution of expression across a panel of ten tissues. The graph shows the number of sequence-verified products that were either not expressed ("0"), expressed in one or more but not all tested tissues ("1" - "9"), and expressed in all tissues tested ("10").

Of 9999 arrayed elements on the two microarrays (including positive and negative controls and "failed" products), 2353 (51%) were expressed in at least one tissue or cell type. Of the gene elements showing significant signal — where expression was scored as "significant" if the normalized Cy3 signal was greater than 1, representing signal 5-fold over biological noise (0.2) — 39% (991) were expressed in all 10 tissues. The next most common class (15%) consisted of gene elements expressed in only a single tissue.

The genes expressed in a single tissue were further analyzed, and the results of the analyses are compiled in FIG. 7.

FIG. 7A is a matrix presenting the expression of all verified sequences that showed expression greater than 3 in at least one tissue. Each clone is represented by a column in the matrix. Each of the 10 tissues assayed is represented by a separate row in the matrix, and relative expression of a clone in that tissue is indicated at the respective node by intensity of green shading, with the intensity legend shown in panel B. The top row of the matrix ("EST Hit") contains "bioinformatic" rather than "physical" expression data — that is, presents the results returned by query of EST, NR and SwissProt databases using the probe sequence. The legend for "bioinformatic expression" (*i.e.*, degree of homology returned) is presented in panel C. Briefly, white is known, black is novel, with gray depicting nonidentical with significant homology (white: E values < 1e-100; gray: E values from 1e-

05 to 1e-99; black: E values > 1e-05).

As FIG. 7 readily shows, heart and brain were demonstrated to have the greatest numbers of genes that were shown to be uniquely expressed in the respective tissue. In brain, 200 uniquely expressed genes were identified; in heart, 150. The remaining tissues gave the following figures for uniquely expressed genes: liver, 100; lung, 70; fetal liver, 150; bone marrow, 75; placenta, 100; HeLa, 50; HBL, 100; and BT474, 50.

It was further observed that there were many more "novel" genes among those that were up-regulated in only one tissue, as compared with those that were down-regulated in only one tissue. In fact, it was found that ORFs whose expression was measurable in only a single of the tested tissues were represented in sequencing databases at a rate of only 11%, whereas 36% of the ORFs whose expression was measurable in 9 of the tissues were present in public databases. As for those ORFs expressed in all ten tissues, fully 45% were present in existing expressed sequence databases. These results are not unexpected, since genes expressed in a greater number of tissues have a higher likelihood of being, and thus of having been, discovered by EST approaches.

Comparison of Signal from Known and Unknown Genes

The normalized signal of the genes found to have high homology to genes present in the GenBank human EST database were compared to the normalized signal of those genes not found in the GenBank human EST database. The data are shown in FIG. 8.

FIG. 8 shows the normalized Cy3 signal intensity for all sequence-verified products with a BLAST Expect ("E") value of greater than 1e-30 (designated "unknown") upon query of existing EST, NR and SwissProt databases, and shows in blue the normalized Cy3 signal intensity for all

sequence-verified products with a BLAST Expect value of less than 1e-30 ("known"). Note that biological background noise has an averaged normalized Cy3 signal intensity of 0.2.

As expected, the most highly expressed of the ORFs were "known" genes. This is not surprising, since very high signal intensity correlates with very commonly-expressed genes, which have a higher likelihood of being found by EST sequence.

However, a significant point is that a large number of even the high expressers were "unknown". Since the genomic approach used to identify genes and to confirm their expression does not bias exons toward either the 3' or 5' end of a gene, many of these high expression genes will not have been detected in an end-sequenced cDNA library.

The significant point is that presence of the gene in an EST database is *not* a prerequisite for incorporation into a genome-derived microarray, and further, that arraying such "unknown" exons can help to assign function to as-yet undiscovered genes.

Verification of Gene Expression

To ascertain the validity of the approach described above to identify genes from raw genomic sequence, expression of two of the probes was assayed using reverse transcriptase polymerase chain reaction (RT PCR) and northern blot analysis.

Two microarray probes were selected on the basis of exon size, prior sequencing success, and tissue-specific gene expression patterns as measured by the microarray experiments. The primers originally used to amplify the two respective ORFs from genomic DNA were used in RT PCR against a panel of tissue-specific cDNAs (Rapid-Scan gene expression panel 24 human cDNAs) (OriGene Technologies,

Inc., Rockville, MD).

Sequence AL079300_1 was shown by microarray hybridization to be present in cardiac tissue, and sequence AL031734_1 was shown by microarray experiment to be present in placental tissue (data not shown). RT-PCR on these two sequences confirmed the tissue-specific gene expression as measured by microarrays, as ascertained by the presence of a correctly sized PCR product from the respective tissue type cDNAs.

Clearly, all microarray results cannot, and indeed should not, be confirmed by independent assay methods, or the high throughput, highly parallel advantages of microarray hybridization assays will be lost. However, in addition to the two RT-PCR results presented above, the observation that 1/3 of the arrayed genes exist in expression databases provides powerful confirmation of the power of our methodology – which combines bioinformatic prediction with expression confirmation using genome-derived single exon microarrays – to identify novel genes from raw genomic data.

To verify that the approach further provides correct characterization of the expression patterns of the identified genes, a detailed analysis was performed of the microarrayed sequences that showed high signal in brain.

For this latter analysis, sequences that showed high (normalized) signal in brain, but which showed very low (normalized) signal (less than 0.5, determined to be biological noise) in all other tissues, were further studied. There were 82 sequences that fit these criteria, approximately 2% of the arrayed elements. The 10 sequences showing the highest signal in brain in microarray hybridizations are detailed in Table 2, along with assigned function, if known or reasonably predicted.

Table 2

| Function of the Most Highly Expressed Genes Expressed Only in Brain | | | | |
|---|---|---|---|---|
| Microarray Sequence Name | Normalized Signal | Expression Ratio | Homology to EST present in GenBank | Gene Function as described by GenBank |
| AP000217-1 | 5.2 | +7.7 | High | S-100 protein, b-chain, $Ca^{2+}$ binding protein expressed in central nervous system |
| AP000047-1 | 2.3 | | High | Unknown Function |
| AC006548-9 | 1.7 | | High | Similar to mouse membrane glyco-protein M6, expressed in central nervous system |
| AC007245-5 | 1.5 | | High | Similar to amphiphysin, a synaptic vesicle-associated protein. Ref 21 |
| L44140-4 | 1.2 | +2.0 | High | Endothelial actin-binding protein found in nonmuscle filamin |

| | | | | |
|---|---|---|---|---|
| AC004689-9 | 1.2 | +3.5 | High | Protein Phosphatase PP2A, neuronal/ downregulates activated protein kinases |
| AL031657-1 | 1.2 | +3.0 | High | Unknown function/ Contains the anhyrin motif, a common protein sequence motif |
| AC009266-2 | 1.1 | +3.7 | Low | Low homology to the Synaptotagmin I protein in rat/present at low levels throughout rat brain |
| AP000086-1 | 1.0 | +2.7 | Low | Unknown, very poor homology to collagen |
| AC004689-3 | 1.0 | | High | Protein Phosphatase PP2A, neuronal/ downregulates activated protein kinases |

Of the ten sequences studied by these latter confirmatory approaches, eight were previously known. Of these eight, six had previously been reported to be important in the central nervous system or brain. The exon

giving the highest signal (AP00217-1) was found to be the gene encoding an S100B $Ca^{2+}$ binding protein, reported in the literature to be highly and uniquely expressed in the central nervous system. Heizmann, *Neurochem. Res.* 9:1097 (1997).

A number of the brain-specific probe sequences (including AC006548-9, AC009266-2) did not have homology to any known human cDNAs in GenBank but did show homology to rat and mouse cDNAs. Sequences AC004689-9 and AC004689-3 were both found to be phosphatases present in neurons (Millward *et al.*, *Trends Biochem. Sci.* 24(5):186-191 (1999)). Two microarray sequences, AP000047-1 and AP000086-1 have unknown function, with AP000086-1 being absent from GenBank. Functionality can now be narrowed down to a role in the central nervous system for both of these genes, showing the power of designing microarrays in this fashion.

Next, the function of the chip sequences with the highest (normalized) signal intensity in brain, regardless of expression in other tissues, was assessed. In this latter analysis, we found expression of many more common genes, since the sequences were not limited to those expressed only in brain. For example, looking at the 20 highest signal intensity spots in brain, 4 were similar to tubulin (AC00807905; AF146191-2; AC007664-4; AF14191-2), 2 were similar to actin (AL035701-2; AL034402-1), and 6 were found to be homologous to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (AL035604-1; Z86090-1; AC006064-L, AC006064-K; AC035604-3; AC006064-L). These genes are often used as controls or housekeeping genes in microarray experiments of all types.

Other interesting genes highly expressed in brain were a ferritin heavy chain protein, which is reported in the literature to be found in brain and liver (Joshi *et al.*, *J. Neurol. Sci.* 134(Suppl):52-56 (1995)), a result

duplicated with the array. Other highly expressed chip sequences included a translation elongation factor 1□ (AC007564-4), a DEAD-box homolog (AL023804-4), and a Y-chromosome RNA-binding motif (Chai *et al.*, *Genomics* 49(2):283-89 (1998))(AC007320-3). A low homology analog (AP00123-1/2) to a gene, DSCR1, thought to be involved in trisomy 21 (Down's syndrome), showed high expression in both brain and heart, in agreement with the literature (Fuentes *et al.*, *Mol. Genet.* 4(10):1935-44 (1995)).

As a further validation of the approach, we selected the BAC AC006064 to be included on the array. This BAC was known to contain the GAPDH gene, and thus could be used as a control for the ORF selection process. The gene finding and exon selection algorithms resulted in choosing 25 exons from BAC AC006064 for spotting onto the array, of which four were drawn from the GAPDH gene. Table 3 shows the comparison of the average expression ratio for the 4 exons from BAC006064 compared with the average expression ratio for 5 different dilutions of a commercially available GAPDH cDNA (Clontech).

Table 3

| Comparison of Expression Ratio, for each tissue, of GAPDH | | |
|---|---|---|
| | AC006064 (n = 4) | Control ( n = 5) |
| Bone Marrow | -1.81 ± 0.11 | -1.85 ± 0.08 |
| Brain | -1.41 ± 0.11 | -1.17 ± 0.05 |
| BT474 | 1.85 ± 0.09 | 1.66 ± 0.12 |
| Fetal Liver | -1.62 ± 0.07 | -1.41 ± 0.05 |
| HBL100 | 1.32 ± 0.05 | 2.64 ± 0.12 |
| Heart | 1.16 ± 0..09 | 1.56 ± 0.10 |
| HeLa | 1.11 ±0.06 | 1.30 ± 0.15 |
| Liver | -1.62 ± 0.22 | -2.07 ± |